Influence du son lors de l'exploration de scènes naturelles dynamiques : prise en compte de l'information sonore dans un modèle d'attention visuelle

Antoine Coutrot

UNIVERSITÉ DE GRENOBLE

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE-ALPES

Spécialité : **Signal, Image, Parole, Télécoms**

Arrêté ministériel : 7 août 2006

Présentée par
Antoine COUTROT

Thèse dirigée par **Alice CAPLIER** et
codirigée par **Nathalie GUYADER**
préparée au sein du
laboratoire Grenoble Image Parole Signal Automatique (Gipsa-lab)
dans **l'école doctorale Electronique Electrotechnique Automatique et Traitement du Signal (EEATS)**

Influence du son lors de l'exploration de scènes naturelles dynamiques
Prise en compte de l'information sonore dans un modèle d'attention visuelle

Thèse soutenue publiquement le **2 octobre 2014**,
devant le jury composé de :

M. Jean-Luc SCHWARTZ
DR CNRS, Grenoble, Président du jury
M. Patrick LE CALLET
Professeur, Université de Nantes, Rapporteur
M. Matei MANCAS
Senior Researcher, Université de Mons, Rapporteur
M. Olivier LE MEUR
MCF, Université de Rennes 1, Examinateur
M. Patrick LAMBERT
Professeur, Université de Savoie, Invité
Mme Alice CAPLIER
Professeur, Université de Grenoble-Alpes, Directrice de thèse
Mme Nathalie GUYADER
MCF, Université de Grenoble-Alpes, Encadrante de thèse

UNIVERSITÉ DE GRENOBLE

ÉCOLE DOCTORALE EEATS

Electronique Electrotechnique Automatique et Traitement du Signal

T H È S E

pour obtenir le titre de

docteur en sciences

de l'Université de Grenoble-Alpes

Mention : Signal, Image, Parole, Télécoms

Présentée et soutenue par

Antoine COUTROT

Influence du son lors de l'exploration de scènes naturelles dynamiques

Prise en compte de l'information sonore dans un modèle d'attention visuelle

Thèse dirigée par Alice CAPLIER et Nathalie GUYADER

préparée au laboratoire Grenoble Image Parole Signal Automatique (Gipsa-lab)

soutenue le 2 octobre 2014

Jury :

Rapporteurs :	M. Patrick LE CALLET	-	Université de Nantes, France
	M. Matei MANCAS	-	Université de Mons, Belgique
Directrice :	Mme Alice CAPLIER	-	Université de Grenoble-Alpes, France
Encadrante :	Mme Nathalie GUYADER	-	Université de Grenoble-Alpes, France
Président :	M. Jean-Luc SCHWARTZ	-	Université de Grenoble-Alpes, France
Examinateur :	M. Olivier LE MEUR	-	Université de Rennes 1, France
Invité :	M. Patrick LAMBERT	-	Université de Savoie, France

Remerciements

Quand je suis arrivé à Grenoble j'avais 20 ans et la ferme intention de devenir astrophysicien.
Sept ans plus tard je suis devenu neuroscientifique après avoir vécu un an au milieu de bateaux bretons et de moutards argentins.

Face à un tel bilan, force est de constater que tout ne s'est pas passé comme prévu.

D'abord, j'ai rencontré de super copains et de non moins super copines, avec lesquels j'ai habité dans le plus surréaliste des appartements (Nathalie Bourgeois, si tu me lis...), mangé les carottes les plus beurrées qui soit, voyagé aux quatre coins de la France (incluant un pique-nique dans le parking à poids lourds du Mont-Saint-Michel), réappris à skier, grimper, marcher.

Puis Anne a sauté à pied joint dans ma vie. Ces quelques lignes ne pouvant ne serait-ce que s'approcher du changement fondamental et du bonheur quotidien que ta présence provoque en moi, je n'en dirai pas plus. Mais n'en penserai pas moins.

Mais tout confit de félicité que j'étais, rien de tout cela n'aurait été possible sans l'ouverture d'esprit de l'équipe pédagogique de PG-Phelma. Pour me laisser faire le grand écart entre la physique statistique et les sciences cognitives en m'offrant un détour d'un an sur les voiliers de l'île d'Arz et les glaciers patagons, il fallait faire preuve de souplesse. Merci.

Puis vint Nathalie, maître de stage d'abord, directrice de thèse ensuite. Ta gentillesse, ta confiance et ta présence rougeoyante m'ont permis de m'épanouir dans ce que j'espère restera mon métier : la recherche scientifique.
Merci Alice pour avoir accepté de chapeauter ces trois années, ton regard frais m'a permis de régulièrement lever la tête du guidon. Merci Gelu pour m'avoir inculqué les bases de la programmation. Merci à l'ensemble des membres du laboratoire avec lesquels j'ai eu de si constructives discussions. Merci tout particulièrement à Jean-Luc. Je suis honoré qu'un chercheur aussi compétent et bienveillant que toi ait accepté de présider mon jury. Merci également à Matei Mancas pour cette profonde semaine à Bilbao, ainsi qu'à Olivier Le Meur, Patrick Le Callet et Patrick Lambert pour avoir évalué mon travail et fait de ma soutenance un si agréable moment.

Je me souviendrai longtemps des innombrables parties de Briscola et de Tressette sans lesquelles les déjeuners n'auraient jamais eu la même saveur. Sans compter la bande Annette-Jonas-Nico-Chloé, dont les parties de Seven, nouvellement bercées par les gazouillis de Maé, resteront parmi les plus chouettes moments de ces

dernières années.

Mais pour pouvoir vivre toute ces aventures, de Grenoble à Lund, de Bahia Blanca au Morbihan, de Toulouse à São Paulo, il est nécessaire d'avoir une solide base d'où décoller et sur laquelle l'on sait pouvoir se replier en toute sérénité. Arthur, François, le 40 rue Boissonade, mon pôpa, ma môman, ma ptite sœur, bref ma famille a toujours pleinement rempli ce rôle, bien au-delà de mes espérances.

Merci.

Table des matières

Glossaire

ascendant	Se dit d'un mécanisme exogène, guidé par un ou plusieurs stimuli extérieurs, indépendamment de la volonté de l'observateur. v, 4
attention	Emprunté du latin *attentio*, "action de tendre son esprit vers". Application de l'esprit, des sens, à un objet, à un fait déterminé. Psychologie : concentration de l'activité mentale sur un point ou sur un objet précis [1]. 4
bas niveau	Se dit d'un traitement codé dans les aires inférieures du cortex. Pour le cortex visuel, il s'agit des régions situées dans le lobe occipital (V1, V2) détectant les contours, les orientations.... vi, 4, 36, 48, 58, 73, 87, 88, 90, 94, 97, 105, 107, 108, 110, 113, 123, 129
bottom-up	voir ascendant. vi, 11, 12, 91
descendant	Se dit d'un mécanisme endogène, lié au vécu, à la volonté, ou à la tâche assignée à l'observateur. vii, 4, 48, 59
fovéa	Zone circulaire de la rétine mesurant environ 1.5 mm, soit 5° dans le champ visuel. Elle présente une densité de photorécepteurs 40 fois plus grande que dans la zone périphérique, et la moitié du cortex visuel lui est dédiée. 4, 6, 13
frame	Une vidéo est constituée d'une succession d'images (25 par seconde), appelées *frames*. 9, 32, 38, 44, 50, 64, 71, 78, 98, 102, 105, 116, 125, 127, 136

1. Le dictionnaire de l'Académie française, 9ème édition, 1992

Gestaltpsychologie — Terme allemand, en français "psychologie de la forme", ou gestaltisme. Théorie psychologique, philosophique et biologique, selon laquelle les processus de la perception et de la représentation mentale traitent spontanément les phénomènes comme des ensembles structurés (les formes) et non comme une simple addition ou juxtaposition d'éléments [2]. 19, 135

haut niveau — Se dit d'un traitement codé dans les aires supérieures du cortex. Pour le cortex visuel, il s'agit des régions détectant les formes (inféro-temporales), les visages (*fusiform face area*)... Ne pas faire d'association trop directe "*bottom-up* = bas niveau" et "*top-down* = haut niveau". Par exemple, la perception des visages est un mécanisme *bottom-up* impliquant des aires visuelles de haut niveau (voir état de l'art sur la perception et l'observation des visages, section 4.1). vi, 4, 36, 48, 58, 74, 88, 90, 105, 108, 113

intégration multimodale — Processus au cours duquel nous transformons un ensemble de stimuli en provenance de différentes modalités en un percept cohérent. 26, 27, 57, 59, 93, 95

psychocinématique — Discipline cherchant à mettre en évidence et comprendre les liens entre les mécanismes cognitifs impliqués dans la perception de films et les intentions du cinéaste [Hasson *et al.* 2008a, Shimamura 2013]. 11, 57, 110

saillance — Propension d'une région ou d'un objet à attirer l'attention. Dans cette thèse, il sera principalement question de saillance *bottom-up*. 13, 17, 18, 21, 28, 31, 35, 48, 49, 56, 59, 92, 97, 99, 105–108, 113, 125, 127, 131, 133, 135, 137, 138

2. `http://en.wikipedia.org/wiki/Gestalt_psychology`

scanpath	Ensemble de saccades et de fixations effectuées par un observateur au cours de l'exploration d'une même scène. 7, 9, 31, 103, 104, 109, 134
top-down	voir descendant. vi, 11, 95
écologique	Stimuli ou conditions expérimentales proches d'une perception naturelle. 4, 6, 24, 30, 35

Acronymes

AIM	Amplitude Instantanée Moyenne. 51
ANOVA	Analyse de la Variance. 41, 55, 66, 79, 102, 116, 130, 175
AV	AudioVisuelle. 41–43, 55
BIC	Critère d'Information Bayésien. 77, 78, 123, 124
CCA	Canonical Correlation Analysis. 135, 136
DER	Diarization Error Rate. 126, 127
DESA	Discrete Energy Separation Algorithm. 49, 53–57, 72
DKL	divergence de Kullback-Leibler. 17, 18, 40, 41, 43, 44, 47, 48, 56, 73, 128–130
EEG	Electro-encéphalographie. 26, 95
EM	Espérance - Maximisation. 75, 79, 82, 176
ETM	Energie de Teager-Kaiser Moyenne. 51
FIM	Fréquence Instantanée Moyenne. 51
fps	frames par seconde. 38, 116
HMM	Hidden Markov Model. 121
IRMf	Imagerie par Résonance Magnétique fonctionnelle. 8, 26
Lasso	Least Absolute Shrinkage and Selection Operator. 76–79, 82, 105, 127, 129, 176
M	Moyenne. 36, 37, 53, 63, 65, 103, 114, 115
MEG	Magnéto-encéphalographie. 26
MFCC	Mel Frequency Cepstral Coefficients. 122–124, 135
NSS	Normalized Scanpath Saliency. 17, 18, 73, 129–131

POM	Plusieurs Objets en Mouvement. 64, 65, 69, 71, 74, 79, 84, 86, 100, 135, 175
ROC	Receiver Operating Characteristic. 17, 73
SD	Standard Deviation (écart-type). 36, 37, 63, 65, 114, 115
SE	Standard Error (erreur standard). 53, 103
SSVEP	Steady-State Visual Evoked Potentials. 95
UOM	Un Objet en Mouvement. 64, 65, 69, 71, 79, 84, 86, 175
V	Visuelle. 41–43, 55
VT	Vérité Terrain. 104

"If the doors of perception were cleansed every thing would appear to man as it is, infinite. For man has closed himself up, till he sees all things through narrow chinks of his cavern."

William Blake, *The Marriage of Heaven and Hell* (1790)

"Ha vous, ne me touchez pas ! Votre vue seule résonne à mes oreilles comme une sirène d'alarme, car dans nos périodes troublées, vous avez toujours été l'étincelle qui fait déborder le vase."

M. le Maire de Champignac

Avant propos

Depuis l'avènement industriel de la fée électricité grâce à laquelle nous pouvons voir même la nuit tombée, la modalité visuelle n'a cessé de prendre de l'importance. Cette domination est d'autant plus forte que les moyens de communication modernes donnent une place de plus en plus importante à l'image. Tablettes, ordinateurs, télévision : les écrans sont omniprésents. Même les téléphones, jusqu'alors l'apanage de la modalité sonore, sont remplacés par les *smartphones*, d'avantage orientés messages textuels, mails et autres visioconférences. Dans son essai *La Société du Spectacle*[3], le philosophe Guy Debord a prédit et critiqué cette hégémonie du médium visuel : « Toute la vie des sociétés dans lesquelles règnent les conditions modernes de production s'annonce comme une immense accumulation de spectacles », définissant le spectacle notamment comme un « rapport social entre des personnes médiatisé par des images ».

Cet âge d'or de la modalité visuelle se retrouve également dans la recherche scientifique, qui l'a longtemps considérée comme la modalité dominante, auto-suffisante et indifférente aux autres sens. Cette conception était cohérente avec une vision modulaire du cerveau, selon laquelle chaque fonction cognitive est compartimentée dans une zone particulière, sans interaction avec les autres régions. Même au sein de la modalité visuelle, un tel cloisonnement était opéré, le traitement de la couleur étant séparé de celui du mouvement, celui de la reconnaissance des formes de celui de la stéréoscopie. A cela venaient s'ajouter les différences physiques objectives entre les signaux issus de nos différents sens, le tout rendant douteux que des entités si différentes puissent interagir.

Cependant, à l'inverse des "singes de la sagesse" dont chacun se cache les yeux, les oreilles ou la bouche, nous percevons bien le monde grâce à tous nos sens à la fois. Si les signaux que nous percevions étaient intrinsèquement incompatibles, quel sens y aurait-il à parler de couleurs chaudes, froides, criardes ou dures, de sons clairs, aigus, éclatants, rocailleux ou moelleux, de bruits mous, de parfums pénétrants[4]. Depuis quelques années, de nombreux champs de recherche, des sciences du langage aux interfaces homme-machine, tentent de comprendre et d'exploiter la nature de ces interactions sensorielles.

Ce manuscrit s'inscrit dans cette démarche, en étudiant tout particulièrement l'influence du son sur l'exploration visuelle de scènes dynamiques.

3. Guy Debord (1967), La Société du Spectacle, Buchet/Chastel.
4. Merleau-Ponty (1947), Le cinéma et la nouvelle psychologie, Les Temps modernes, vol. 3, no. 26, pages 930-947.

Chapitre 1

Etat de l'art

"Attention is like water. It flows. It's liquid. You create channels to divert it, and you hope that it flows the right way."

Apollo Robbins, *gentleman thief*[1]

Sommaire

Le monde dans lequel nous évoluons est si riche et si complexe que notre cerveau ne peut l'apprécier dans sa globalité. Afin d'optimiser le traitement perceptif des informations nous parvenant, nous utilisons un processus appelé l'attention. Ce mécanisme permet de filtrer les informations les plus pertinentes afin de leur allouer un maximum de ressources cognitives, au détriment des autres stimuli. Dans un premier temps, nous présenterons une sélection des principaux résultats comportementaux de la littérature ayant trait à l'attention visuelle, puis auditive (nous n'aborderons

1. `https://www.youtube.com/watch?v=ib-5m3SyJeA`. Voir aussi [Otero-Millan *et al.* 2011]

pas les nombreux travaux effectués sur le sujet en neuroimagerie). Nous verrons ensuite que ces deux types d'attention sont étroitement liés et interagissent fortement. Enfin, nous passerons en revue certaines des nombreuses applications liées à cet immense champ de recherche, et situerons nos travaux au sein de ce dernier.

1.1 Vocabulaire

Mais avant cela, il nous semble important de préciser quelques notions de vocabulaire, afin de partir sur des bases communes. En effet, les nombreux domaines d'étude utilisant ces concepts sont souvent issus de cultures différentes, et une définition précise pour les uns peut s'avérer inexacte pour les autres. Nous encourageons donc le lecteur à se reporter au glossaire, notamment pour les mots attention, ascendant, descendant, haut niveau, bas niveau, écologique...

1.2 Attention visuelle

L'attention visuelle fut la première à avoir été étudiée. D'abord parce qu'elle a longtemps été considérée comme la modalité dominante, liée à la notion même de conscience. Ensuite parce que les aires visuelles sont les plus étendues de notre cerveau (20 à 25 % du cortex [Wandell *et al.* 2007]). Enfin parce que contrairement aux oreilles pour l'audition, les yeux s'orientent le plus souvent vers l'objet d'intérêt, en permettant une mesure directe. En effet, au centre de notre rétine se trouve la zone d'acuité maximale : la *fovéa*. Cette zone circulaire mesure environ 1.5 mm, soit 5 ° dans le champ visuel. Elle présente une densité de photorécepteurs 40 fois plus grande que dans la zone périphérique de la rétine et la moitié du cortex visuel lui est dédiée. Comme on peut le constater Figure 1.1, il est parfaitement naturel de placer les zones que l'on souhaite analyser en détail au centre de la fovéa, et donc de notre œil, afin d'y allouer les meilleures ressources de notre système visuel. Cette caractéristique implique un lien étroit entre attention et mouvements oculaires. Cependant, ce lien n'est pas toujours évident. Il existe deux mécanismes distincts d'allocation attentionnelle : l'attention *overt* (flagrante) et *covert* (cachée) [Hoffman 1998]. On parle d'attention *covert* lorsque l'on porte mentalement notre attention vers une région particulière du champ visuel, sans y déplacer les yeux. A l'inverse, l'attention *overt* se traduit par le déplacement des yeux vers la région d'intérêt. La théorie prémotrice de l'attention introduite par Rizzolati et collègues [Rizzolati *et al.* 1987] propose que l'allocation attentionnelle *overt* et les mouvements oculaires partagent les mêmes bases neuronales : l'attention provoquerait la planification d'une saccade vers la région concernée. Cette théorie s'est trouvée renforcée par de récentes expériences neurophysiologiques qui ont montré qu'une stimulation intracraniale (sous le seuil de déclenchement) de certaines structures oculomotrices amélioraient la sensibilité

Figure 1.1 – Exemple d'image avec un filtrage spatialement variant modélisant la distribution non uniforme des photorécepteurs sur la rétine. Ici, l'œil regarde la zone de l'image marquée par une flèche à environ 57cm de distance. Extrait de [Séré *et al.* 2000]

Paramètres	**Ordres de grandeur**
Durée d'une fixation	200 - 400 ms
Durée d'une saccade	10 - 60 ms
Latence d'une saccade	200 - 250 ms
Amplitude d'une saccade	1 - 10°
Durée d'une microsaccade	6 - 15 ms
Amplitude d'une microsaccade	0.5 - 1°

Table 1.1 – Ordres de grandeur caractéristiques des principaux mouvements oculaires lors de l'exploration libre de scènes naturelles.

visuelle de la région rétinotopique correspondante [Belopolsky & Theeuwes 2009]. Bien que d'autres théories plaident pour une plus grande séparation de ces deux fonctions [Klein 1980], l'existence d'une grande corrélation entre les mouvements oculaires et l'attention visuelle fait néanmoins consensus. Etudier les mouvements oculaires revient donc à ouvrir une fenêtre sur le fonctionnement de bien des mécanismes cognitifs en lien avec nos processus attentionnels.

1.2.1 Mouvements oculaires

Depuis 1959 et la publication "What the frog's eye tell the frog's brain" [Lettvin *et al.* 1959], nous savons que l'œil joue un rôle important dans la sélection et l'organisation de l'information visuelle. En effet, l'œil a deux fonctions : la *perception*, i.e. le codage de l'information lumineuse, et *l'action motrice*, i.e. amener le regard sur les régions d'intérêt, celles qui ont attiré notre attention. Si les yeux d'une grenouille sont assez statiques, ceux des primates ne cessent de bouger, volontairement ou de

manière réflexe. Il existe un nombre limité de mouvements oculaires, les principaux étant les saccades, les poursuites continues, les fixations, et les microsaccades. Leurs ordres de grandeur lors de l'exploration de scènes naturelles sont consignés Table 1.1.
- La saccade est un mouvement rapide des deux yeux dans la même direction. Il s'agit certainement du mouvement le plus rapide dont nous sommes capables, puisqu'il peut atteindre 900 degrés angulaire par seconde, sur une durée très brève [Rayner 1998]. Les saccades sont caractérisées par leur point de départ, leur amplitude (en degré angulaire), leur vitesse (en degré par seconde) et leur direction. Leur rôle est de placer une région du champ visuel au niveau de la fovéa. Les saccades sont des mouvements très stéréotypés. Leur profil de vitesse présente une seule accélération suivie d'une seule décélération, ce qui en facilite la détection automatique.
- La poursuite continue (ou *smooth pursuit*) se manifeste lors de la poursuite d'un stimulus en mouvement, pour le garder au centre de la fovéa. Plus lent que les saccades, ce mouvement a généralement une vitesse d'une centaine de degrés par seconde.
- La fin d'une saccade ou d'une poursuite appelle le début d'une fixation, c'est-à-dire la stabilisation des yeux sur une position du champ visuel le temps de percevoir l'information qu'elle contient. Une fixation se caractérise par sa position spatiale et sa durée.
- Même au cours d'une fixation, nos yeux continuent de bouger, sans que nous en ayons conscience : ce sont les microsaccades.

Les mouvements oculaires étant d'une part une fenêtre ouvrant en temps réel sur de nombreux processus cognitifs, et d'autre part plus faciles d'accès qu'une exploration directe et intrusive du cerveau, ils sont depuis longtemps l'objet d'intenses recherches. Dans un premier temps, les chercheurs ont étudié les mouvements oculaires de personnes regardant des stimuli artificiels ou des scènes naturelles statiques. Puis, les connaissances et les techniques évoluant, ils ont cherché à utiliser des stimuli de plus en plus écologiques, comme des vidéos. Actuellement, de plus en plus d'études utilisent des oculomètres "tête libre" montés sur des lunettes et permettant au participant de se déplacer dans l'espace.

1.2.2 Exploration de scènes naturelles

Les bases des recherches sur l'exploration de scènes *via* les mouvements oculaires furent jetées dès les années 1920 et quelques études, désormais devenues des classiques, furent menées durant la première moitié du XX$^{\text{ème}}$ siècle. En 1935, Buswell observa les séquences de mouvements oculaires de sujets regardant une série d'œuvres d'art représentant des scènes complexes (voir Figure 1.2) [Buswell 1935]. Il constata que les différents sujets avaient un comportement assez similaire tant en termes de position oculaire que de distribution des durées de fixation. En 1967, Yar-

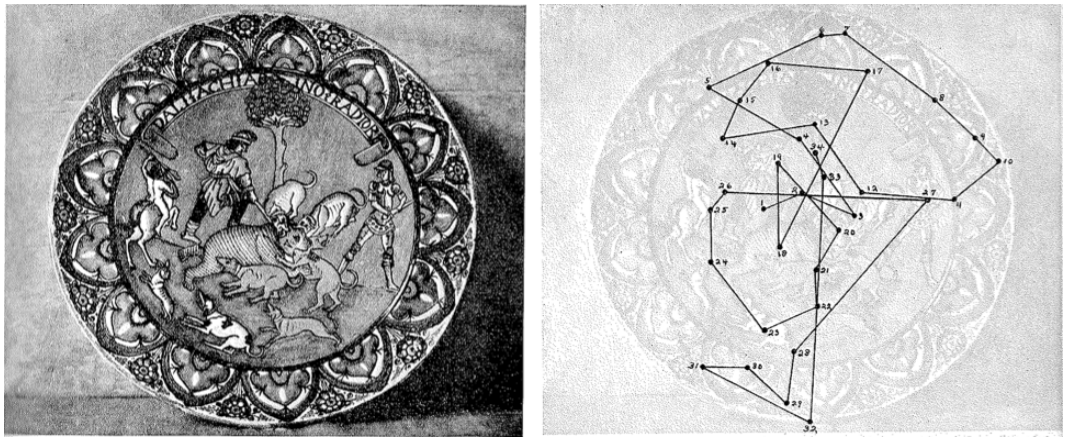


Figure 1.2 – Saccades (traits) et fixations (points) enregistrés par Buswell lors de son expérience au moyen d'un appareil fonctionnant selon le même principe que les oculomètres modernes. A droite, les fixations du *scanpath* sont numérotées par ordre chronologique. Extrait de [Buswell 1935].

bus fut l'un des premiers à démontrer la nature cognitive des mouvements oculaires. Il enregistra les saccades et fixations de personnes auxquelles était assignée une tâche particulière et constata que les séquences de mouvements oculaires variaient de manière systématique avec la tâche demandée [Yarbus 1967]. Mais c'est après 1970 et les innovations technologiques permettant une mesure simple et précise des mouvements oculaires que ce domaine s'est réellement développé.
Depuis, de nombreuses études ont mis en évidence différents facteurs influençant l'exploration visuelle de scènes naturelles. En Figure 1.3, nous représentons certains de ces facteurs en les classant selon leur nature : ascendante, descendante, ou les deux à la fois. Dans la suite de cette section nous allons très brièvement passer en revue les facteurs descendants, ces derniers n'étant pas au cœur de notre travail. Puis, nous nous intéresserons plus en détail aux processus ascendants que nous croiserons régulièrement au cours de ce manuscrit.

1.2.2.1 Facteurs descendants

- La tâche Le stimulus le plus connu de l'histoire de l'oculométrie est sans doute la toile "Retour Inattendu" (1888) du peintre russe Ilya Repin, grâce auquel Yarbus a montré que les scanpaths [2] changeaient du tout au tout en fonction de la tâche assignée aux observateurs [Yarbus 1967]. Depuis, ce résultat a été répliqué lors de nombreuses études modernes, lesquelles ont par exemple montré que les observateurs effectuent davantage de fixations sur les objets présents dans la scène lors d'une tâche de mémorisation comparé à une tâche de recherche [Henderson & Hollingworth 1998, Castelhano *et al.* 2009, Mills *et al.* 2011, Smith & Mital 2013].

2. voir glossaire

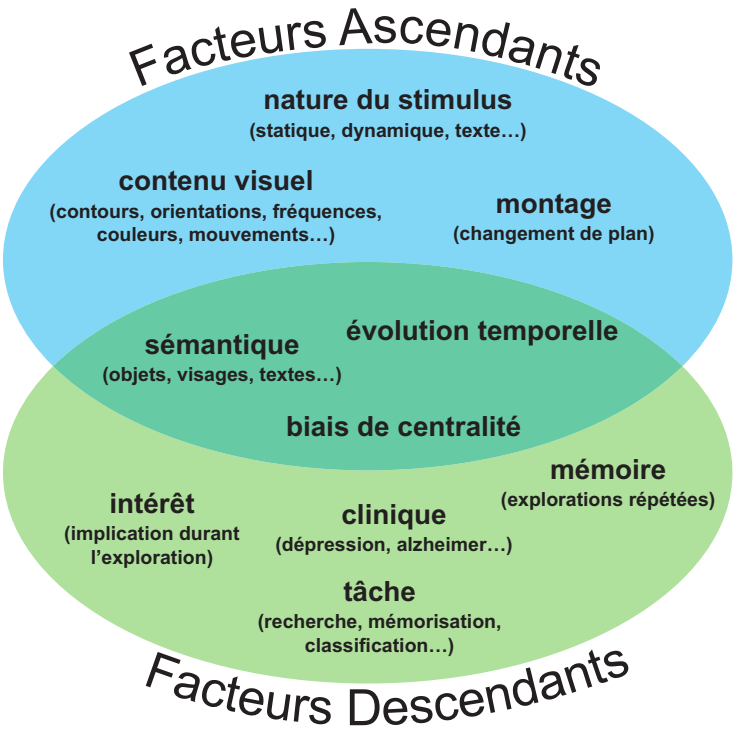


Figure 1.3 – Facteurs ascendants et descendants influençant l'exploration visuelle.

- L'intérêt Il parait légitime de penser que si un stimulus vous passionne, vous le regarderez différemment que s'il vous ennuie. Pour tester cette hypothèse, une équipe a mesuré les mouvements oculaires ainsi que l'activité cérébrale (obtenue par Imagerie par Résonance Magnétique fonctionnelle (IRMf)) de participants regardant un extrait vidéo réalisé par un professionnel et reconnu comme un chef-d'œuvre [3], un extrait de film professionnel de bonne qualité [4], un extrait d'une série télévisée [5], et une vidéo prise par une caméra fixe dans un parc [Hasson *et al.* 2008a]. Les auteurs ont constaté une gradation dans la corrélation des activités cérébrales inter-sujets selon la qualité cinématographique du stimulus (65% du cortex des participants présentaient une réponse similaire pour le film d'Hitchcock, contre 45% pour celui de Sergio Leone, 18% pour la série télévisée, et moins de 5% pour la vidéo sur caméra fixe). Ce résultat signifie que le contrôle esthétique du contenu des stimuli joue un rôle important dans l'engagement des observateurs dans leur perception.

- Mémoire Nous n'explorons pas de la même manière une scène la première fois ou après des explorations répétées [Foulsham & Underwood 2008, Harding &

3. Alfred Hitchcock, *Bang ! You're Dead* (1961)
4. Sergio Leone, *The Good, the Bad and the Ugly* (1966)
5. Larry David, *Curb Your Enthusiasm* (2000)

Bloj 2010, Hadizadeh *et al.* 2012]. La similarité entre les scanpaths des différents participants diminue avec le nombre de présentations. Cependant, après un jour sans présentation du stimulus, la similarité revient à son niveau initial, puis recommence à baisser avec la répétition des présentations [Dorr *et al.* 2010].

- Clinique Certaines pathologies (Alzheimer, schizophrénie) ont un effet direct sur le contrôle oculaire. Ainsi des recherches utilisant l'oculométrie sur des paradigmes de génération de "prosaccades" (saccades réflexes vers une cible périphérique), et d'"antisaccades" (saccades volontaires à l'opposé d'une cible périphérique nécessitant en amont l'inhibition de la saccade réflexe), permettent d'étudier, de mieux comprendre, voir de diagnostiquer certaines de ces pathologies [Marendaz *et al.* 2007].

1.2.2.2 Facteurs ascendants

- Nature du stimulus Nous n'explorons pas de la même façon un stimulus s'il s'agit d'une image statique ou d'une scène dynamique. Nous effectuons des fixations plus longues, des saccades plus grandes et la variance entre les positions oculaires de différents participants est plus faible sur des vidéos de scènes naturelles que sur images statiques [Smith & Mital 2013].
Cependant, ces résultats sont à prendre avec précaution dans la mesure où ils peuvent être fortement influencés par d'autres paramètres, comme la tâche ou le contenu visuel des stimuli. Ce point sera plus longuement discuté au troisième chapitre, section 3.4.

- Montage La dynamique temporelle des stimuli visuels influence drastiquement leur exploration. Dorr *et al.* l'ont systématiquement manipulée en comparant les mouvements oculaires enregistrés sur des films "naturels", sur des bandes-annonces de films commerciaux (*Star Wars*, *War of the Worlds*), sur des films en "stop-motion" (les films naturels ont été temporellement sous-échantillonnés à une frame toutes les 90), et sur des images statiques (toujours issues des films naturels) [Dorr *et al.* 2010]. Les auteurs montrent que les observateurs effectuent davantage de grandes ($>10°$) et de petites ($<5°$) saccades sur les films naturels que dans les autres conditions. La cohérence entre les positions oculaires des différents sujets augmente brutalement dès que le contenu visuel est renouvelé (changement de plan pour les bandes-annonces, changement d'image pour le stop-motion), les observateurs se rapprochant du centre de la scène à son apparition, avant de s'en éloigner progressivement (voir ci-dessous le point "biais de centralité").
Les relations entre changement de plan et regard des spectateurs sont si fortes que certains modèles sont explicitement bâtis sur leur rapport l'un à l'autre. Ainsi, dans [Boccignone *et al.* 2005], les auteurs proposent un modèle de détection des

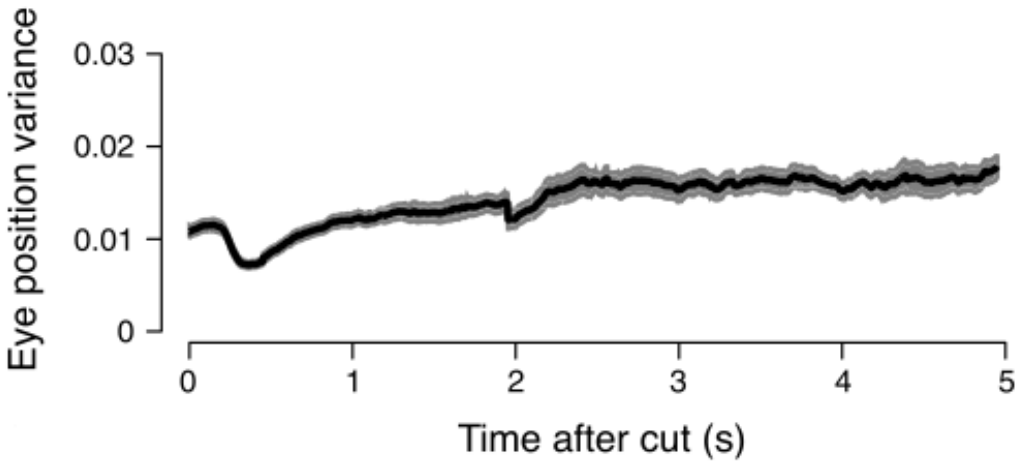
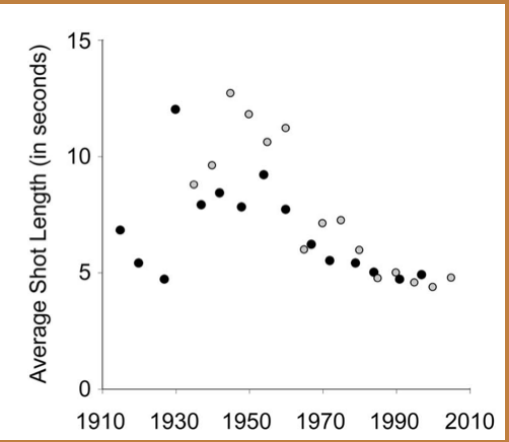


Figure 1.4 – Variance entre les positions oculaires de 10 observateurs durant les 5 secondes suivant un changement de plan. Les ombres entourant la courbe correspondent à l'erreur standard. Extrait de [Wang *et al.* 2012].

changements de plans basé sur l'analyse de la variance des positions oculaires d'un observateur idéal regardant une vidéo. Les changements de plans induisant une discontinuité dans l'exploration visuelle de la scène, les auteurs les détectent en les identifiant aux irrégularités d'une fonction de cohérence entre les positions oculaires en fonction du temps. A l'inverse, dans [Wang *et al.* 2012], les auteurs utilisent les changements de plans dans un modèle de prédiction des mouvements oculaires sur des vidéos. Les auteurs proposent un modèle assez simple : au début de chaque plan, les observateurs cherchent, localisent et suivent du regard les zones d'intérêt de la scène, chaque changement de plan réinitialisant le processus. Ce modèle donne de très bons résultats sans prendre en compte d'informations haut niveau tel le contexte narratif dans lequel le plan s'inscrit. En effet, les auteurs ont tenté d'éliminer ce dernier en découpant leurs stimuli en plans de différentes durées et en modifiant leur ordre d'apparition. Malgré ces profondes modifications, les auteurs ont systématiquement obtenu la même évolution de variance entre les positions oculaires au cours d'un plan, voir Figure 1.4. La diminution de la variance quelques millisecondes après chaque changement de plan correspond à l'augmentation de la cohérence constatée dans [Dorr *et al.* 2010].

- Contenu visuel Afin de comprendre ce qui, dans une image, attire le plus notre attention, de nombreuses études ont tenté d'établir un lien entre les zones de fixation et les propriétés physiques (couleur, orientation, contraste...) de ces régions. Une des premières études parues sur le sujet fut celle de Mannan et collègues [Mannan *et al.* 1995] qui établit que les régions fixées ne diffèrent pas significativement lorsque l'on filtre les hautes ou basses fréquences de l'image. Les auteurs en déduisent que le regard doit être attiré par des propriétés locales de l'image peu affectées par ces modifications globales. Ils mettent en évidence une corrélation entre lieux de fixation et

Psychocinématique 1



Evolution de la durée moyenne des plans sur deux bases de plus de 13 000 films parus entre 1910 et 2010. Extrait de [Cutting *et al.* 2011].

Les premiers films de l'histoire du cinéma n'étaient pour la plupart composés que d'un seul plan fixe. Rapidement, les cinéastes se mirent à combiner différents plans afin de rendre le film plus riche, dynamique et proche de la vision naturelle. Certains auteurs ont même comparé ces coupures à notre mode d'exploration visuelle saccadée [Wagner *et al.* 2006]. Au fil des années, l'évolution des techniques de tournage, de mise en scène et de montage a créé des règles destinées à mieux capter l'attention du spectateur, en lissant le plus possible ces discontinuités et en établissant un lien logique entre les différents plans. Ces règles, connues sous le nom de *continuity editing rules* ou *Hollywood style* [Smith 2012, Shimamura 2013], sont abondamment utilisées par certains cinéastes. Les réalisateurs utiliseraient (consciemment ou non) la baisse de variance entre les positions oculaires provoquée par les changements de plans pour mieux guider l'attention des spectateurs [DeLong *et al.* 2012]. Ceci pourrait expliquer la baisse continue de la durée des plans illustrée sur la figure ci-dessus, environ 65% en un siècle.

maxima locaux de contraste et de densité de contour [Mannan *et al.* 1996]. Ces résultats corroborent ceux d'une étude menée par Tatler *et al.* dans laquelle les auteurs comparent systématiquement les zones fixées avec toute une gamme d'attributs locaux : la luminance, la chromaticité, le contraste et la densité de contour [Tatler *et al.* 2005]. Chaque attribut est testé pour 13 échelles spatiales différentes. Leurs résultats indiquent que chacun de ces attributs joue un rôle dans la sélection attentionnelle, surtout pour les hautes fréquences spatiales, et à plus forte raison pour le contraste et la densité de contour.

Dans [Carmi & Itti 2006], Carmi *et al.* tentent de déterminer l'impact des attributs dynamiques sur les mouvements oculaires *bottom-up*. Pour distinguer ces derniers de ceux liés à des processus *top-down*, les auteurs ne travaillent qu'avec les saccades survenant dans les 250 ms suivant un changement de plan, et conduisant à une forte cohérence entre les participants. La Figure 1.5 montre que les attributs dynamiques (contraste de clignotement, de mouvement) sont bien plus efficaces pour prédire

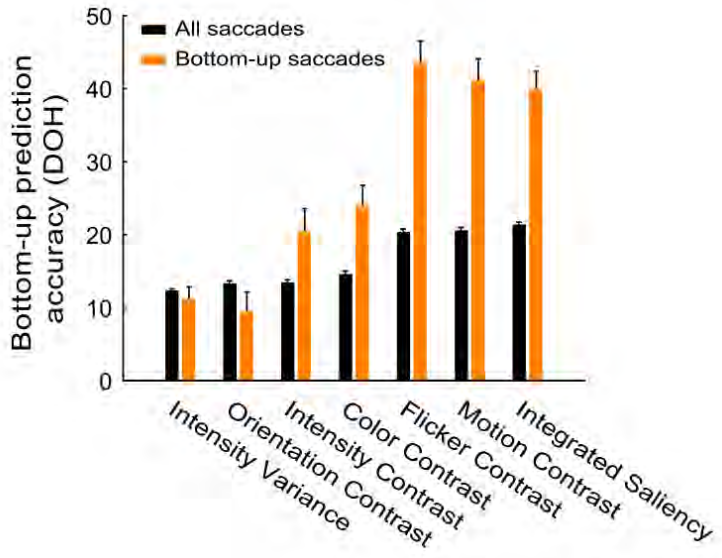


Figure 1.5 – Pouvoir prédictif de différents attributs de "toutes les saccades" versus "les "saccades *bottom-up*" (DOH = différence d'histogrammes). Une saccade est dite *bottom-up* si elle a été effectuée dans les 250 ms suivant un changement de plan et si elle a conduit à une forte cohérence des positions oculaires entre les différents sujets. Extrait de [Carmi & Itti 2006].

les lieux de fixations que les attributs statiques (variation d'intensité, contraste d'orientation, de couleur, d'intensité), surtout pour les mouvements oculaires *bottom-up*. Cette prépondérance des attributs dynamiques pour guider le regard se retrouve dans de nombreuses études [Tosi *et al.* 1997, Goldstein *et al.* 2007, Dorr *et al.* 2010, Mital *et al.* 2010, Smith & Mital 2013] et n'est guère surclassée que par certains attributs haut niveau comme les visages. Mais nous reviendrons en détail sur ce point au quatrième chapitre.

1.2.2.3 Facteurs mixtes

- Evolution temporelle Certaines études suggèrent que les stratégies d'exploration varient considérablement avec le temps. Il existerait deux phases d'exploration visuelle : une phase de découverte de la scène (courtes fixations et grandes saccades), suivie d'une phase d'exploration plus détaillée (longues fixations et petites saccades) [Buswell 1935, Antes 1974, Unema *et al.* 2005, Velichkovsky *et al.* 2005, Pannasch *et al.* 2008, Mills *et al.* 2011]. Cependant, cette idée est sujet à controverse, et d'autres études mettent plutôt en évidence une alternance entre plusieurs phases de découverte et d'exploration [Follet *et al.* 2011]. Comme pour l'influence de la tâche, l'évolution temporelle de l'exploration visuelle pourrait interagir avec de multiples facteurs et mener à des résultats différents selon leur combinaison. Nous reviendrons sur ce point section 3.4.1.2.

- Biais de centralité Le biais central d'exploration traduit la tendance à regarder davantage le centre d'une image que les régions périphériques, surtout durant les premières millisecondes suivant l'apparition du stimulus. Ce biais a été beaucoup étudié car il affecte toutes les expériences d'oculométrie utilisant des scènes naturelles [Buswell 1935, Tosi *et al.* 1997, Parkhurst *et al.* 2002, Tatler 2007, Tseng *et al.* 2009, Dorr *et al.* 2010, Gautier & Le Meur 2012, Marat *et al.* 2013, Smith 2013, Smith & Mital 2013]. Ce biais, prédominant durant les premières fixations suivant l'apparition du stimulus, s'explique si l'on considère que le centre de l'écran est le lieu optimal pour saisir rapidement l'essence de la scène, repérer ses zones d'intérêt avant de commencer l'exploration à proprement parler. En effet, comme l'acuité visuelle décline lorsque l'on s'éloigne de la fovéa, commencer par fixer une position à la périphérie de la scène entraînerait une mauvaise résolution, et donc une perte d'information sur le côté opposé du stimulus. Plusieurs explications (ascendantes et descendantes) au biais central ont été proposées :

- *Le biais du photographe* : les images de scènes naturelles utilisées dans les expérience d'oculométrie sont souvent extraites de bases de photos prises par des professionnels, lesquels tendent à placer les objets les plus intéressants au centre de l'image.
- *Le biais moteur* : il est moins coûteux de faire de courtes saccades. Or, dans les expériences d'oculométrie, les stimuli sont souvent précédés d'une croix centrale de fixation, concentrant dès le début les regards au centre de l'écran.
- *La stratégie d'exploration* : cette explication est liée au biais du photographe. Comme nous sommes habitués à voir les objets d'intérêt au centre de l'image, nous sommes naturellement enclins à regarder cette position. De plus, si aucun objet d'intérêt n'est présent dans la scène, le centre est l'endroit le plus stratégique pour en guetter l'apparition.
- *La réserve orbitale* : la position de repos des yeux est le centre.
- *Le présentation centrale* : Les stimuli sont présentés au centre d'un écran, lui même placé au centre du champ visuel des sujets.

Dans [Tseng *et al.* 2009], les auteurs montrent que le biais central est principalement lié au biais du photographe ainsi qu'à la stratégie d'exploration lors de l'apparition du stimulus. Le biais moteur, la réserve orbitale et la présentation centrale y contribuent dans une moindre mesure.

1.2.3 Modélisation

Etant donnés le nombre et la diversité des facteurs susceptibles de l'influencer, modéliser l'exploration visuelle est une gageure. De très nombreux modèles de saillance ont été proposés et parviennent, plus ou moins bien selon le contexte, à prédire les zones les plus susceptibles d'attirer l'attention. Le lecteur peut se reporter à [Borji & Itti 2013], qui dresse une taxonomie de près de 65 modèles. Dans

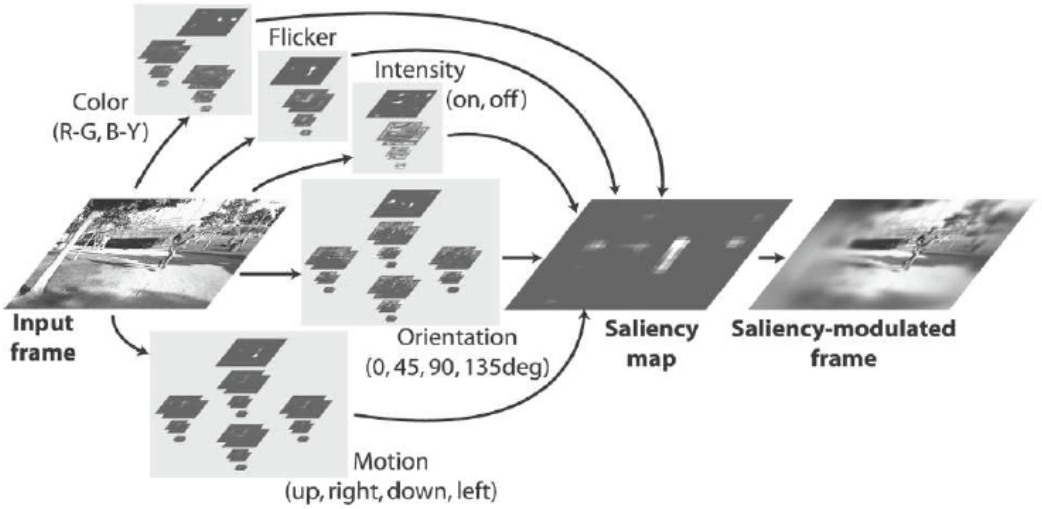
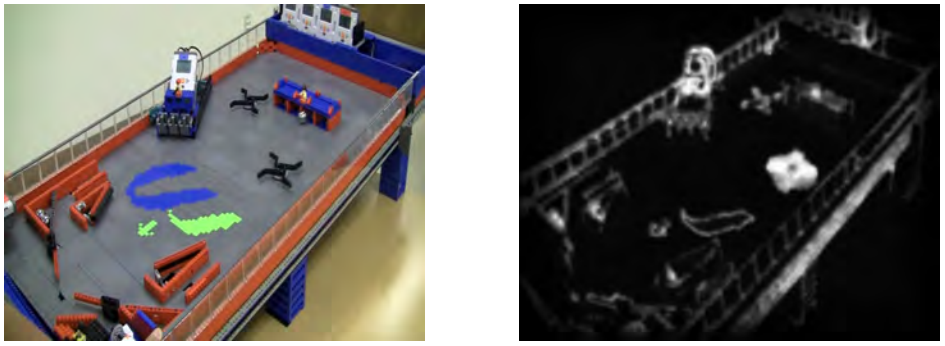


Figure 1.6 – Vue d'ensemble d'un modèle de saillance visuelle. Pour chaque frame, le modèle crée un jeu de cartes multi-échelles par attribut statique (orientation, couleur, intensité) et dynamique (mouvements compensés du mouvement de caméra, clignotements). Les attributs sont ensuite mis en compétition, normalisés puis fusionnés en une carte de saillance maîtresse. Extrait de [Itti 2005].

cette section, nous nous contenterons de présenter leurs fondements théoriques, et décrirons plus en détail le modèle proposé par Marat *et al.* dont nous servirons régulièrement au cours de cette thèse [Marat *et al.* 2009]. Enfin, nous présenterons les principales métriques utilisées pour évaluer leurs performances.

1.2.3.1 Origines

La plupart des modèles de saillance se basent sur la la *Feature Integration Theory* (FIT), introduite dans [Treisman & Gelade 1980]. La FIT fait l'hypothèse que l'information visuelle est décomposée dans le cortex en cartes d'attributs élémentaires distinctes (orientation, contraste, couleur...), traitées en parallèle sur tout le champ visuel. Les différentes cartes d'attribut sont ensuite associées en une carte de saillance maîtresse (la *master saliency map*), dont les zones les plus activées seront les plus saillantes, celles vers lesquelles l'attention exogène se déploiera prioritairement.
Le modèle d'Itti, Koch et Niebur est l'une des premières implémentations computationnelles de cette théorie et est devenu une référence [Itti *et al.* 1998]. Il décompose d'abord l'entrée visuelle (une image ou une frame de vidéo) en un jeu de cartes topographiques multi-échelles correspondant à chaque attribut élémentaire (intensité, chrominance et orientation). Plus tard ont été ajoutés des attributs dynamiques, comme le mouvement (Figure 1.6). Les différentes positions spatiales sont mises en compétition de manière à ce qu'au sein de chaque carte, seules les



(a) Image originale. **(b)** Carte de saillance associée.

Figure 1.7 – L'intensité lumineuse de la carte de droite est proportionnelle à la saillance de l'image de gauche. La roue de droite du flipper est particulièrement saillante car en mouvement. Cartes calculées à partir du modèle décrit dans [Marat *et al.* 2009].

positions les plus contrastées par rapport à leur voisinage soient mises en valeur. Toutes les cartes sont ensuite normalisées et fusionnées en une carte de saillance maîtresse, comme l'illustre la Figure 1.7.

1.2.3.2 Modèle de Marat *et al.*

Il s'agit d'un modèle ascendant, biologiquement inspiré, et basé sur l'information de luminance. Initialement constitué d'une voie statique et d'une voie dynamique [Marat *et al.* 2009], il a par la suite été amélioré avec l'ajout d'une troisième voie incluant un détecteur de visages [Marat *et al.* 2013], voir Figure 1.8. Tout d'abord, un algorithme de compensation du mouvement dominant est appliqué afin d'éliminer d'éventuels mouvements de caméra. Puis deux étapes se succèdent : une représentant le fonctionnement de la rétine, l'autre celui des cellules complexes de l'aire V1 du cortex visuel primaire. L'étape rétinienne ne modélise pas la distribution de photorécepteurs sur la rétine mais extrait d'une part les basses fréquences spatiales pour la voie dynamique (sortie "magnocellulaire" de la rétine), et d'autre part les hautes fréquences spatiales pour la voie statique (sortie "parvocellulaire" de la rétine). L'étape corticale utilise ensuite un banc de filtres de Gabor pour séparer les images filtrées selon 6 orientations et 4 bandes fréquentielles.

Voie statique Les sorties des filtres de Gabor sont normalisées pour favoriser les images ayant des maxima spatialement distribués. Elles sont ensuite fusionnées en une carte de saillance statique faisant ressortir les zones ayant de forts contrastes de luminance.

Voie dynamique Sous l'hypothèse de conservation du flux optique, le mouvement est estimé pour chaque sortie des filtres de Gabor. Puis, un filtre temporel médian est appliqué afin de lisser chaque carte, lesquelles sont fusionnées en une

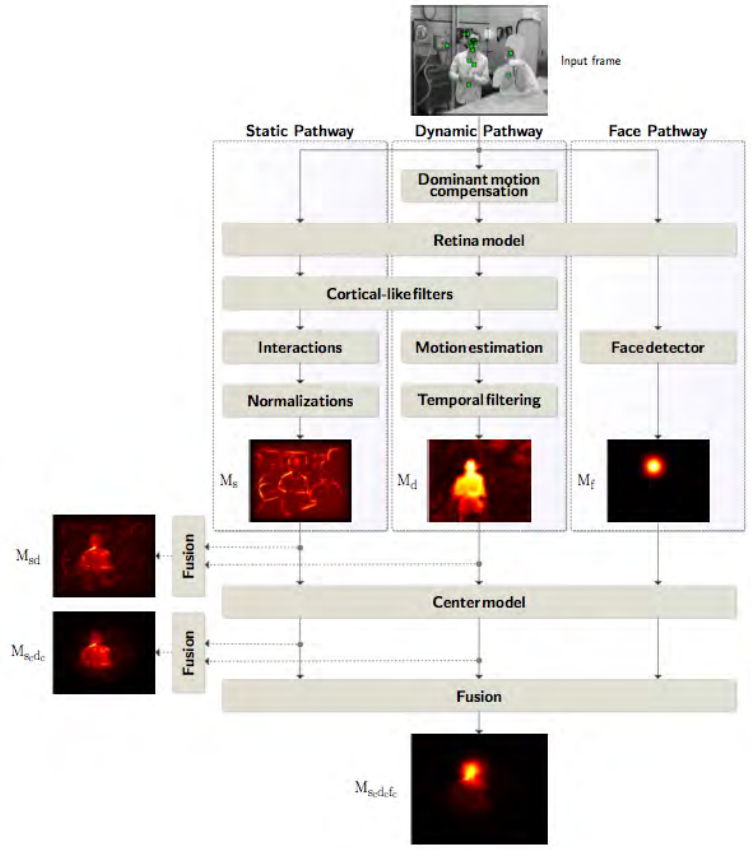


Figure 1.8 – Architecture du modèle proposé par Marat *et al.*. En entrée, l'image est séparée en trois voies : statique, dynamique et visage. Les sorties de ces voies passent ensuite par un modèle de biais de centralité, avant d'être fusionnées en une carte de saillance maîtresse. Extrait de [Marat *et al.* 2013].

carte de saillance dynamique. Cette carte met en valeur les zones en mouvement, affichant son amplitude.

Voie visage Les visages sont détectés au moyen d'un détecteur type Viola-Jones [Viola & Jones 2004]. Une gaussienne 2D, dont les dimensions dépendent de l'indice de confiance de détection, est ensuite utilisée pour souligner la saillance de chaque visage.

Ces trois sorties sont ensuite fusionnées en une carte de saillance maîtresse. Une telle opération n'a rien de trivial car elle rassemble des cartes de natures et d'amplitudes différentes. De nombreuses approches ont été proposées, allant de la simple moyenne à l'apprentissage par noyaux multiples [Zhao & Koch 2012, Kavak *et al.* 2013]. Pour une revue des différentes techniques, se reporter à [Chamaret *et al.* 2010, Muddamsetty *et al.* 2013].

Marat *et al.* sont partis du principe qu'une frame est saillante si sa carte statique a un fort maximum global, et si sa carte dynamique ne présente que quelques régions en mouvement. Ainsi, ils ont choisi de pondérer la carte statique M_s par son maximum

global, et la carte dynamique M_d par le coefficient d'asymétrie de sa distribution (*skewness*). La carte visage M_v est quant à elle pondérée par l'indice de confiance de détection. De plus, hypothèse est faite que les régions de saillance communes à plusieurs cartes doivent être renforcées, introduisant des termes non-linéaires. Ils expriment donc la carte de saillance maitresse M comme suit :

$$M = \alpha M_s + \beta M_d + \gamma M_v + \alpha\beta M_s M_d + \beta\gamma M_d M_v + \alpha\gamma M_s M_v \tag{1.1}$$

avec $\alpha = \max(M_s), \beta = \text{skewness}(M_d)$ et $\gamma = \text{mean}(\text{confidence})$.

1.2.3.3 Evaluation de la saillance visuelle

L'évaluation des performances d'un modèle de saillance visuelle se fait en comparant les zones prédites par ce dernier avec celles effectivement regardées par des observateurs lors d'une expérience oculométrique. Plusieurs métriques ont été définies pour mener à bien cette tâche, voir [Le Meur & Baccino 2013] pour une revue complète. Dans [Riche *et al.* 2013b], les classements de 12 modèles de saillance ont été comparés au moyen de 12 métriques différentes. Ceci a permis de mettre en évidence une redondance entre certaines métriques, alors que d'autres mènent effectivement à des ordres de classement différents. Les auteurs recommandent donc l'usage de trois métriques qui à elles seules permettent d'évaluer les performances d'un modèle dans leur ensemble. Il s'agit du Normalized Scanpath Saliency (NSS) [Peters *et al.* 2005], de la divergence de Kullback-Leibler (DKL) [Kullback & Leibler 1951], et de l'aire sous la courbe ROC proposée dans [Borji *et al.* 2012]. Cette dernière étant plutôt conçue pour détecter la saillance d'objets de manière binaire (1 saillant, 0 non saillant), elle ne sera pas adaptée à l'évaluation de carte de saillance dont les valeurs s'échelonnent de manière continue, comme c'est le cas du modèle de Marat *et al.* Aussi, nous n'utiliserons dans ce manuscrit que le NSS et la DKL, dont voici les définitions.

• La DKL mesure la distance entre deux probabilités de distribution. Cette métrique peut être comparée à une mesure de corrélation pondérée entre deux fonctions de densité de probabilité. Pour évaluer un modèle, on calcule pour chaque frame la DKL entre la carte de saillance prédite S et la carte des positions oculaires enregistrées O.
La carte O est construite en sommant une gaussienne 2D d'écart-type $\sigma = 1°$, centrée sur chacune des n positions oculaires $(x_i, y_i)_{i\in[1..n]}$ enregistrées sur la frame considérée (une illustration est disponible au chapitre suivant, Figure 2.2).

$$O(x, y) = \sum_{j=1}^{n} \exp - \left(\frac{(x - x_j)^2}{2\sigma^2} + \frac{(y - y_j)^2}{2\sigma^2} \right) \tag{1.2}$$

Les deux cartes S et O sont ensuite normalisées de manière à constituer des densités de probabilité. la DKL symétrique entre ces deux cartes (de p pixels) est définie de la manière suivante.

$$DKL(S,O) = \frac{1}{2}\left(\sum_{i=1}^{p} S_i \log\frac{S_i}{O_i} + \sum_{i=1}^{p} O_i \log\frac{O_i}{S_i}\right) \tag{1.3}$$

Plus la DKL est grande, plus les deux cartes sont différentes, et donc moins le modèle est bon. Une DKL nulle correspond à une prédiction parfaite.

• Le NSS utilise les valeurs de la carte de saillance S prédite par le modèle prises aux n positions oculaires enregistrées. La carte S est centrée-réduite afin de diminuer la valeur du NSS si son écart-type σ_s est important ou si toutes ses valeurs sont proches d'une moyenne μ_s. En effet, une carte de saillance sera d'autant plus précise qu'elle fera ressortir un nombre restreint de régions d'intérêt, et que ses valeurs aux positions oculaires seront grandes par rapport à la moyenne. Comme nous ne regardons rarement qu'un pixel en particulier, il est d'usage d'appliquer une petite gaussienne 2D centrée sur chaque position oculaire comme nous venons de le voir pour la DKL.

$$S^{cr} = \frac{S-\mu_s}{\sigma_s} \text{ et } NSS = \frac{1}{np}\sum_{i=1}^{p} S_i^{cr}.O_i \tag{1.4}$$

avec "." l'opérateur multiplication terme à terme et p le nombre de pixels des cartes S et O. Le NSS est proportionnel à la pertinence du modèle : plus il est positif (respectivement négatif), plus les positions oculaires et les régions prédites sont corrélées (respectivement anticorrélées). Il n'est pas borné.

Les performances des modèles de saillance varient beaucoup avec le type de stimuli, la tâche assignée à l'observateur (ou sa volonté), ou encore avec le contexte dans lequel la scène s'inscrit. Certains modèles ont proposé d'intégrer ces paramètres dans des modèles de saillance descendants [Milanese *et al.* 1994, Torralba *et al.* 2006, Tsotsos *et al.* 2008, Borji *et al.* 2011, Xu *et al.* 2014]. Néanmoins, ces modèles ont toujours un domaine d'application assez restreint et le défi que constitue un modèle universel prenant en compte le ou les objectifs sous-tendant toute exploration n'a pas encore été relevé [Tatler *et al.* 2011].

L'attention visuelle en général et l'oculométrie en particulier ont donc fait l'objet d'intenses recherches depuis près d'un siècle, et de nombreux facteurs permettant de modéliser la façon dont nous explorons visuellement notre environnement ont été identifiés. Ceci est moins vrai pour l'attention auditive qui, moins facile à mesurer, a moins été étudiée. Cependant, certains remarquables travaux ont permis de lever une partie du voile posé sur ce domaine de recherche.

1.3 Attention auditive

J'entend des voix. Lueurs à travers ma paupière.
Une cloche est en branle à l'église Saint-Pierre.
Cri des baigneurs. Plus près ! plus loin ! non, par ici !
Non, par là ! Les oiseaux gazouillent. Jeanne aussi.
George l'appelle. Chant des coqs. Une truelle
Racle un toit. Des chevaux passent dans la ruelle.
Grincement d'une faulx qui coupe le gazon.
Chocs. Rumeurs. Des couvreurs marchent sur la maison.
Bruit du port. Sifflement des machines chauffées.
Musique militaire arrivant par bouffées.
Brouhaha sur le quai. Voix françaises. Merci.
Bonjour. Adieu. Sans doute il est tard, car voici
Que vient tout près de moi chanter mon rouge-gorge.
Vacarmes de marteaux lointains dans une forge.
L'eau clapote. On entend haleter un steamer.
Une mouche entre. Souffle immense de la mer.

Victor Hugo, *Le matin. - En dormant.*
Dans la section Fenêtres Ouvertes de *L'art d'être grand-père* (1877)

1.3.1 Analyse de scènes auditives

Les sons sont des objets pouvant se représenter sur un trièdre temps-fréquence-intensité. Le temps caractérise le début, la fin ou les modulations, la fréquence caractérise le timbre et la hauteur, et l'intensité caractérise l'énergie contenue dans le signal. Depuis le début du XX$^{\text{ème}}$ siècle et la *Gestaltpsychologie*[6], nous savons que notre cerveau transforme les contours, couleurs et formes perçues individuellement par notre rétine en objets unitaires cohérents. De la même façon, il transforme les modulations temporelles et fréquentielles perçues par notre tympan et cochlée en flux sonores cohérents.
Cette transformation, baptisée "analyse de scènes auditives" a été théorisée par Albert Bregman en 1990 [Bregman 1990]. Elle repose sur deux opérations : le groupement (*grouping*) et la séparation (*streaming*). Si différents attributs élémentaires sont groupés, nous les attribuons à une même source ou "flux" sonore. S'ils sont séparés, nous les répartissons entre plusieurs sources. Deux mécanismes ascendants permettent, étant donné un ensemble d'attributs, ou "traits acoustiques", de décider s'ils doivent être groupés ou séparés.

Traits de groupement simultané

6. voir glossaire

- Harmonicité (périodicité des stimuli).
- Cohérence d'enveloppe (les stimuli démarrent ensemble)
- Modulation cohérente de fréquence (les spectrogrammes des stimuli ont même allure)
- Corrélation binaurale (les stimuli sont perçus de la même façon par les deux oreilles)

Traits de groupement séquentiel

- Similarité fréquentielle (une succession de stimuli de fréquences proches)
- Similarité de timbre (idem pour des stimuli de timbres proches)
- Répétition (une succession de stimuli à un certain rythme)

Ces mécanismes de groupement ascendants sont automatiques. Il existe aussi des situations où le choix groupement / séparation se base sur des "schémas" descendants. Par exemple, si dans un flux de parole, on remplace un phonème par un bruit assez fort, le phonème absent est perçu comme présent, comme si le flux de parole avait continué à exister derrière le bruit. Un autre exemple célèbre est celui de l'effet *Cocktail Party* [Cherry 1953]. Cet effet opère lorsque, dans une salle de réception baignant dans un brouhaha continue, nous arrivons en concentrant notre attention sur notre interlocuteur à séparer le flux de parole provenant de sa bouche de ceux des autres convives.

1.3.2 L'ouïe : précise en temps et en fréquence, moins en espace

L'oreille humaine est capable de reconnaître précisément un large spectre de fréquences avec une excellente résolution temporelle (entre 2 et 10 ms de délai de conduction de l'oreille au cerveau contre environ 50 ms de l'œil au cerveau). Cependant, elle est nettement moins performante lorsqu'il s'agit de localisation spatiale, ne possédant pas d'organe, équivalent à la rétine pour l'œil, lui permettant de dresser directement une carte topographique.
De la même manière qu'il se sert de la disparité binoculaire pour extraire l'information de profondeur, notre cerveau se sert de la différence de temps que met un son pour arriver d'une oreille à l'autre pour estimer sa source. Deux autres grandeurs physiques aident à la localisation des sources sonores : la différence interaurale de niveau (différence d'intensité du son entre la première et la seconde oreille à laquelle il parvient), et la différence interaurale de phase (idem pour la phase). L'audition humaine fait intervenir d'autres paramètres plus complexes : déformation des fréquences par le pavillon des oreilles variant suivant leur direction, conduction osseuse... L'ensemble de ces facteurs peut être modélisé au moyen d'une "fonction de transfert relative à la tête" (HRTF, *head-related transfer function*).
Il faut enfin mentionner les mécanismes descendants concourant à la localisation spatiale : on aura davantage tendance à chercher la source d'un son de moteur sous le capot d'une voiture que dans les arbres. Cependant, l'oreille humaine a une résolution spatiale très moyenne comparée à celle de l'œil : si ces deux résolutions

spatiales étaient identiques, nous serions capables de repérer à 15 m un déplacement de 5 cm d'un instrument de musique [7], ce qui compliquerait diablement la tâche des chefs d'orchestre.

1.3.3 Modélisation

Le premier modèle de saillance sonore ne datant que de 2005, de nombreux défis restent à relever. Il existe deux classes de modèles.
Les premiers tentent d'estimer la localisation spatiale des sources sonores grâce aux indices physiques précédemment décrits. La plupart de ces modèles ont été conçus pour l'orientation de robots humanoïdes [Ruesch *et al.* 2008, Schauerte *et al.* 2011, Deleforge & Horaud 2012, Zaraki *et al.* 2014].
Les seconds s'appuient sur les similitudes entre le système visuel et le système auditif et ont une structure identique aux modèles de saillance visuelle tels que nous venons de présenter [Kayser *et al.* 2005, Tsuchida & Cottrell 2012, Kaya & Elhilali 2012]. Ils ont été conçus pour de multiples applications comme la détection de parole dans des environnements bruités [Evangelopoulos & Maragos 2006], la détection des syllabes accentuées dans un flux de parole [Kalinli & Narayanan 2007], la détection d'attaques dans les chants [Coath *et al.* 2007], l'analyse de scènes auditives complexes [Duangudom & Anderson 2007] ou encore le design sonore [De Coensel & Botteldooren 2010]. C'est avec cette classe de modèles que nous travaillerons dans ce manuscrit. Leur architecture générale est la suivante. Tout d'abord, le signal sonore est converti en une carte d'intensité temps-fréquence par transformée de Fourier (étape cochléaire). Cette carte est ensuite considérée comme une image à laquelle sont appliqués les mêmes traitements que ceux des modèles de saillance visuelle (étape corticale). Différents attributs élémentaires sont extraits à différentes échelles : intensité, contraste fréquentiel, contraste temporel, orientations, hauteur... Enfin, ils sont mis en compétition, normalisés et fusionnés en une carte de saillance temps-fréquence, comme l'illustre la Figure 1.9. Certains auteurs réalisent une moyenne fréquentielle afin de n'avoir plus qu'une courbe de saillance en fonction du temps. Cette représentation peut s'avérer plus adaptée pour repérer les événements sonores d'un signal [Zlatintsi *et al.* 2012].

Une nouvelle et prometteuse classe de modèles divergeant de cette architecture "classique" est en train de voir le jour. Il s'agit de modèles bayésiens basés sur le concept de "surprise" [Schauerte & Stiefelhagen 2013]. Dans cette approche, un modèle probabiliste de la distribution des fréquences dans le signal mesure le degré de "surprise" perçue par un auditeur en fonction de ses précédentes observations. Ces modèles ont l'avantage d'être moins gourmands en temps de calcul et de ne pas avoir besoin d'information sur le futur pour estimer la saillance du présent, comme le nécessitent l'analyse multi-échelle et le filtrage opérés par les modèles précédents.

7. www.snv.jussieu.fr/vie/dossiers/auditionvision

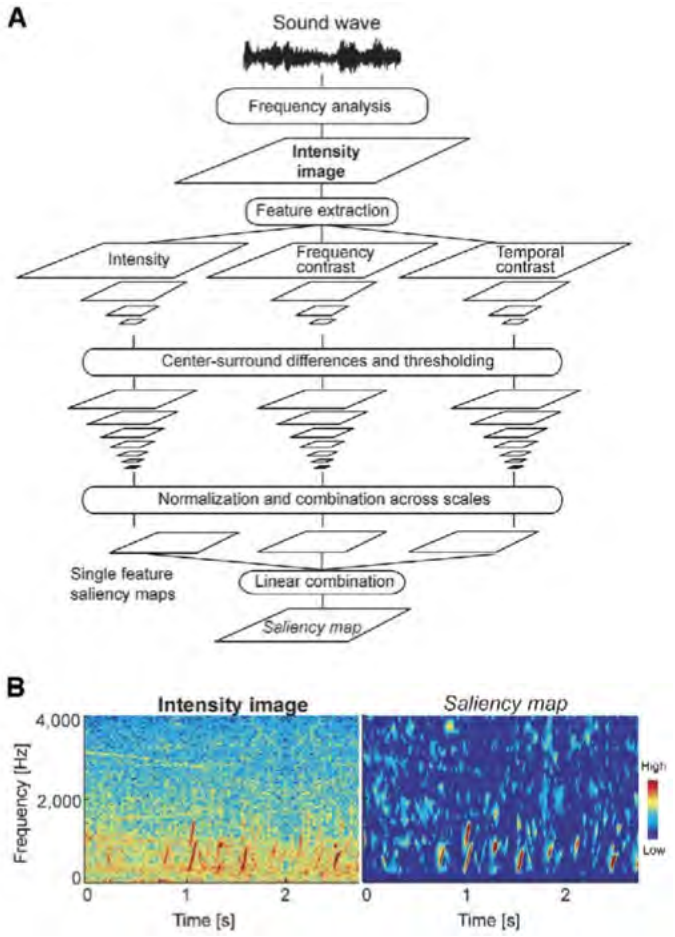


Figure 1.9 – (**A**) - Vue d'ensemble du modèle de Kayser *et al.* (**B**) - Spectrogramme de bulles d'eau dans un environnement bruité (gauche) et sa carte de saillance (droite). Extrait de [Kayser *et al.* 2005].

Ils sont ainsi mieux adaptés à de potentielles problématiques "temps-réel".

1.3.4 Evaluation de la saillance sonore

Contrairement à la saillance visuelle que l'oculométrie rend relativement aisée à évaluer, il n'existe pas de corrélat physique facilement mesurable et reflétant l'attention auditive. En d'autres termes, nous bougeons rarement les oreilles. Pour contourner cette difficulté, les chercheurs ont redoublé d'inventivité et ont proposé plusieurs méthodes, plus ou moins efficaces.
- Dans [Kayser *et al.* 2005], les auteurs ont diffusé 150 paires de stimuli sonores de 1.2 s sur un fond de bruit blanc ou de scène sonore aléatoire. Les participants devaient choisir pour chaque paire lequel des deux stimuli était le plus saillant. L'expérience a été menée avec des humains et avec des macaques. Ces derniers

étaient assis entre 2 haut-parleurs diffusant chacun un stimulus, et l'expérimentateur regardait vers lequel s'orientait le regard de l'animal.
- Dans [Kalinli & Narayanan 2007], les auteurs ont annoté manuellement les mots et les syllabes accentués de leurs stimuli et les ont comparé aux prédictions de leur modèle.
- Dans [Coath *et al.* 2007], les auteurs ont fait la même chose avec les changements abrupts en énergie et en fréquence de leurs stimuli.
- Dans [De Coensel & Botteldooren 2010], les auteurs ont demandé à 20 groupes de 5 participants (avec des quotas d'âge, de genre et de sensibilité auditive) de se réunir dans une maison dans laquelle ils pouvaient vaquer à leurs occupations. A l'extérieur de la maison étaient installés des haut-parleurs émettant un fond sonore de trafic routier ou de trains passant à différentes distances. La tâche des participants était de d'estimer à quel point un stimulus les avait gêné, relativement au stimulus précédent.
- Dans [Tsuchida & Cottrell 2012], les auteurs ont utilisé le même paradigme que Kaiser *et al.*, en demandant cette fois aux participants de juger quel stimulus était le plus "intéressant". Le but était d'avoir une estimation de l'intérêt de leurs stimuli, qu'ils pensent être monotonement lié à la saillance sonore.
- Dans [Kaya & Elhilali 2012], les auteurs ont caché, dans un fond sonore, une cible auditive dont ils ont fait varier le timbre, la hauteur et intensité. La performance de leur modèle était proportionnelle à sa capacité à retrouver cette dernière.
- Enfin, Varinthira Duangudom Delmotte a soutenu une thèse sur la saillance sonore computationnelle, dans laquelle elle propose en détail une méthode d'évaluation particulièrement intéressante [Duangudom Delmotte 2012]. Les participants devaient réaliser deux tâches simultanément. La tâche primaire consistait à compter le nombre de tons basse fréquence présents dans un flux audio composé de sons de 200 ms à 100 Hz et 200 Hz séparés d'un intervalle de 2 ms. La tâche secondaire consistait à repérer la présence d'une cible (son d'amplitude modulée) au sein de 4 distracteurs (sons à 570 Hz, 700 Hz, 840 Hz et 1000 Hz). Les sons des deux tâches sont spectralement disjoints pour éviter tout phénomène de masquage. Les auteurs font alors l'hypothèse que la performance des participants sur la seconde tâche est proportionnelle à la saillance de la cible.

L'ensemble de ces méthodes, si intéressantes soient-elles, souffre du même défaut : elles n'ont été conçues que pour évaluer la saillance d'un son ponctuel, isolé de son contexte. Or il paraît légitime de penser que la saillance d'un son dépend grandement de ceux qui le précèdent, voir qui lui succèdent. A la section 2.2.2, nous proposons une évaluation de la saillance sonore permettant de tenir compte du contexte général.

	Signal	
Différences	Visuel	Sonore
Support physique	onde électromagnétique	onde acoustique
Vitesse de propagation	300 000 km/s	0.340 km/s
Entrée sensorielle	rétine	tympan + cochlée
Délai de conduction	50 ms	2-10 ms
Perception des fréquences	mauvaise	bonne
Localisation spatiale	bonne	mauvaise

Table 1.2 – Différences principales entre les signaux visuel et sonore.

1.4 Attention audiovisuelle

"Sitting in my Study I hear a Coach drive along the street ; I look through the Casement and see it ; I walk out and enter into it ; thus, common Speech would incline one to think, I heard, saw, and touch'd the same thing, to wit, the Coach. It is nevertheless certain, the Ideas intromitted by each Sense are widely different, and distinct from each other ; but having been observed constantly to go together, they are spoken of as one and the same thing."

Bishop George Berkeley, *Essay Towards a New Theory of Vision* (1732)

"Quand tu sais te faire entendre, ne t'étonne pas d'être mal vu."

Bernard Lubat, musicien de jazz

Malgré les différences fondamentales séparant le son de l'image (Table 1.2), les cloisons académiques séparant hermétiquement nos sens se fissurent, et les interactions intermodales deviennent un sujet d'étude privilégié. Les illusions audiovisuelles comme la ventriloquie ou l'effet McGurk (sur lesquelles nous reviendront au chapitre 4, section 4.1.2.1) comptent parmi les manifestations les plus spectaculaires de l'intégration du son avec l'image. Cependant, s'il est aisé de les constater, les comprendre l'est nettement moins.
Dans un premier temps, nous allons décrire les contraintes bas niveau nécessaires à l'intégration de deux stimuli simples. Puis, nous ferons un état de l'art des études ayant utilisé des scènes plus complexes et écologiques, en nous attardant particulièrement sur celles se servant des mouvements oculaires comme indice de l'intégration. Enfin, nous décrirons quelques intéressantes applications utilisant certains de ces résultats.

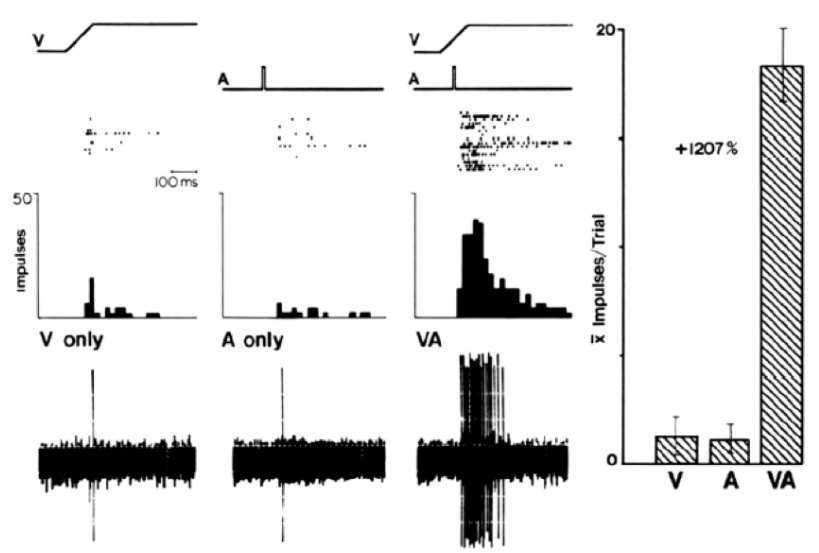


Figure 1.10 – Réponse de neurones du coliculus supérieur de chats à des stimuli bimodaux visuo-auditif (VA), unimodaux visuels (V) ou unimodaux auditifs (A). Du haut vers le bas : décours temporel des stimuli, réponses par essai (chaque point est un potentiel d'action, chaque ligne est un essai), histogrammes et oscillogrammes représentatifs. La moyenne du nombre de potentiels d'action dans le cas bimodal est augmentée de 1207% par rapport aux cas unimodaux. Extrait de [Meredith & Stein 1986].

1.4.1 Fondations

Les premières études établissant de manière quantitative l'existence d'une intégration audiovisuelle datent des années 1980. Meredith & Stein ont enregistré au moyen d'électrodes intracrâniennes la réponse électrique de neurones du coliculus supérieur (CS) de chats auxquels étaient présentés des stimuli visuels, sonores, ou une combinaison des deux [Meredith & Stein 1986, Stein & Meredith 1993]. Les auteurs ont constaté (voir Figure 1.10) que les réponses des neurones du CS à des stimuli simultanés bimodaux sont bien plus fortes que la somme des réponses aux stimuli unimodaux (de 1207%), ce qui témoignent bien d'un processus d'intégration des deux modalités. En faisant varier l'intensité ainsi que la synchronisation spatiale et temporelle des stimuli, les auteurs ont énoncé trois règles :

Règle spatiale La réponse d'un neurone multimodal est maximale lorsque les stimuli multisensoriels coïncident spatialement. Cette réponse est souvent supérieure à la somme des réponses unimodales. Inversement, lorsque les stimuli sont spatialement disparates, il y a absence ou dépression de la réponse du neurone, alors inférieure aux réponses unimodales. Cette augmentation n'a pas lieu pour deux stimuli de la même modalité [Kadunce *et al.* 1997].

Règle temporelle L'amplitude de la réponse multimodale décroit avec le temps séparant les deux stimuli candidats à l'intégration. Cependant, la fenêtre temporelle au sein de laquelle l'intégration est possible est relativement large (jus-

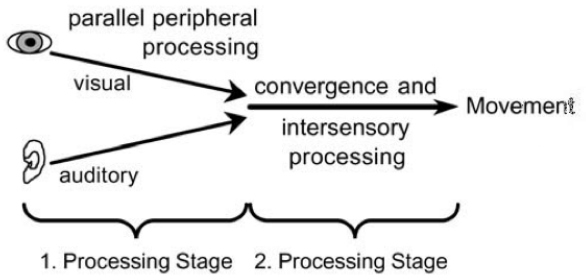


Figure 1.11 – Modèle en deux étapes. Dans un premier temps, les stimuli visuel et auditif sont traités séparément. Dans un second temps, ils convergent et sont intégrés. Extrait de [Arndt & Colonius 2003]

qu'à 1500 ms). Cette propriété permet d'intégrer des stimuli qui, bien qu'issus de la même source, n'arrivent pas en même temps à nos capteurs sensoriels du fait de leur différence de vitesse de conduction dans l'espace et dans le système nerveux (voir Table 1.2).

Règle d'efficacité inverse Plus les stimuli unimodaux sont faibles, plus leur combinaison produit un gain important (par gain, les auteurs entendent le pourcentage de gain apporté par la combinaison de stimuli unimodaux par rapport à la réponse maximale à des stimuli unimodaux seuls).

Si la mesure directe de la réponse neuronale à des stimuli bimodaux *via* des électrodes intracrâniennes est une méthode efficace pour mettre en évidence l'existence d'une intégration multimodale bas niveau chez l'animal, ce paradigme est, pour des raisons éthiques, nettement plus compliqué à appliquer chez l'homme. Les chercheurs ont donc contourné ce problème en menant d'une part des expériences utilisant des techniques d'exploration cérébrale moins invasives (EEG, MEG, IRMf...), et d'autre part des paradigmes comportementaux. L'inconvénient de tels paradigmes est qu'il est souvent difficile de contrôler si l'effet observé est effectivement le fruit d'une intégration bas niveau, ou si au contraire il est la conséquence de mécanismes plus haut niveaux.

1.4.2 Stimuli simples

1.4.2.1 La prosaccade "audiovisuelle"

La prosaccade est un paradigme classique pour étudier les processus d'intégration audiovisuelle. Le sujet est face à un écran, le regard maintenu sur un point de fixation central. Une cible apparaît d'un côté ou de l'autre du point de fixation, et le sujet a pour consigne de la fixer le plus rapidement possible. La latence de la saccade, c'est-à-dire le temps séparant l'apparition de la cible du début du mouvement oculaire, est un précieux indice pour comprendre la façon dont le

cerveau traite l'information.
Conformément aux règles spatiales et temporelles énoncées par Meredith & Stein, les latences de réaction à des stimuli audiovisuels présentés proches dans l'espace et dans le temps sont plus courtes que celles de réactions à des stimuli unimodaux [Engelken & Stevens 1989, Perrott *et al.* 1990, Hughes *et al.* 1994, Nozawa *et al.* 1994, Frens *et al.* 1995, Goldring *et al.* 1996, Corneil & Munoz 1996, Corneil *et al.* 2002]. Ceci suggère que l'intégration multimodale joue un rôle important dans la construction des réactions motrices appropriées [Stein & Meredith 1993].

Une des premières explications à cette réduction des latences fut basée sur la facilitation statistique [Raab 1962]. Cette idée propose que le déclenchement de la saccade est une réaction soit au stimulus visuel soit au stimulus auditif, hypothèse faite que ces stimuli sont traités indépendamment et que leurs latences se recouvrent. Les modèles basés sur cette idée appartiennent à la famille des *race models* : le processus sensorimoteur traité le premier sera celui à l'origine de la saccade. La latence prévue par ce modèle devrait donc être la plus petite des latences des stimuli unimodaux [Miller 1982]. Mais dans les faits, les temps de réaction mesurés sont inférieurs à ceux prédits par les *race models* [Hughes *et al.* 1994, Nozawa *et al.* 1994]. Ceci met en défaut les hypothèses énoncées plus haut et suggère que les informations issues de modalités différentes (visuelle et auditive) ne sont pas traitées indépendamment l'une de l'autre mais sont bien intégrées dans le cerveau.

Dans [Arndt & Colonius 2003], les auteurs mesurent une diminution de la latence des saccades vers une cible visuelle lorsque son apparition est synchronisée avec un signal sonore. Cette diminution est d'autant plus prononcée que le signal sonore est fort et proche de la cible visuelle. Néanmoins, les auteurs n'ont pas mesuré d'interaction significative entre l'intensité du signal sonore et sa distance avec la cible visuelle, ce qui laisse à penser que l'intensité du signal sonore n'a pas d'influence directe sur l'intégration audiovisuelle et que les informations liées à l'intensité et à la position spatiale sont traitées à des étapes différentes, comme l'illustre la Figure 1.11. La première étape traite les signaux unimodaux, et peut donc être affectée par le changement de paramètres unimodaux, comme l'intensité, alors que la seconde étape est bimodale et peut donc être affectée par le changement de paramètres bimodaux, comme la distance relative entre les deux stimuli.

1.4.2.2 La spatialisation n'est pas nécessaire à l'intégration

"La musique n'existe que dans le temps sans le moindre rapport à l'espace"

Arthur Schopenhauer, *Le Monde comme Volonté et comme Représentation* (1819)

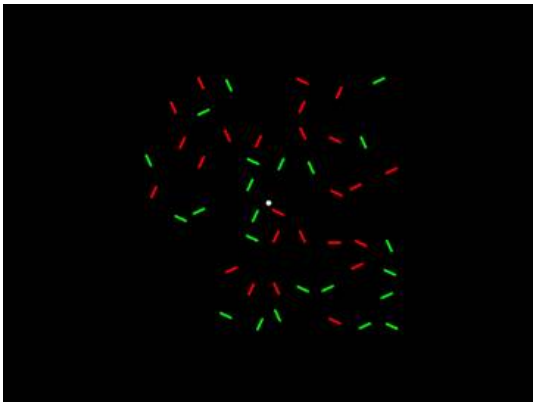


Figure 1.12 – Paradigme du *Pip and Pop phenomenon*. La cible est un segment vertical ou horizontal parmi un certain nombre de distracteurs. A des intervalles de temps aléatoires, un nombre aléatoire de segments change de couleur. Lorsque la cible change de couleur, c'est l'unique segment à le faire. Si un son (*pip*) monophonique est joué en même temps qu'un changement de couleur de la cible, le temps de détection de cette dernière est drastiquement réduit : elle *pop* hors de son environnement. Extrait de [Van der Burg *et al.* 2008].

Dans les travaux que nous avons jusqu'ici présentés, la proximité spatiale des stimuli visuel et sonore semble jouer un rôle déterminant dans leur intégration. Cependant, de nombreuses études ont depuis mis en évidence l'intégration d'un son non spatialisé (monophonique) avec un stimulus visuel [Vroomen & de Gelder 2000, Doyle & Snowden 2001, Olivers & Van der Burg 2008, Chen & Yeh 2009]. En particulier, une spectaculaire étude a montré qu'un signal sonore non spatialisé peut drastiquement réduire le temps de recherche d'une cible visuelle synchronisée avec ce dernier [Van der Burg *et al.* 2008]. Dans cette étude, les participants devaient chercher un segment vertical ou horizontal (cible) parmi 24, 36 ou 48 segments obliques (distracteurs). A des intervalles de temps aléatoires, un nombre aléatoire de segments changeait de couleur. Lorsque la cible changeait de couleur, c'était l'unique segment à le faire (voir Figure 1.12). Les auteurs ont logiquement mesuré une croissance linéaire du temps de réponse avec le nombre de distracteurs. Par contre, lorsqu'un signal sonore monophonique était synchronisé avec le changement de la cible visuelle, les temps de réponse diminuaient drastiquement, et indépendamment du nombre de distracteurs. Afin de contrôler que cet effet était bien le fruit d'une intégration audiovisuelle des deux signaux et non pas un simple effet d'alerte provoqué par le signal sonore, les auteurs ont refait l'expérience en synchronisant le changement de couleur de la cible avec la disparition temporaire (60 ms) d'un signal visuel (un disque ou un halo coloré). La synchronisation de cet indice visuel avec le changement de la cible n'a en rien amélioré sa détection, ce qui infirme l'hypothèse du signal d'alerte. Après d'autres contrôles que nous passons ici sous silence, les auteurs concluent que leur étude met en évidence un liage audiovisuel automatique entre une cible visuelle et un signal sonore non spatialisé. Cet effet a été baptisé le *Pip and Pop phenomenon* : la saillance du *pip* sonore s'intègre avec la saillance de la cible visuelle, provoquant le *pop* de cette

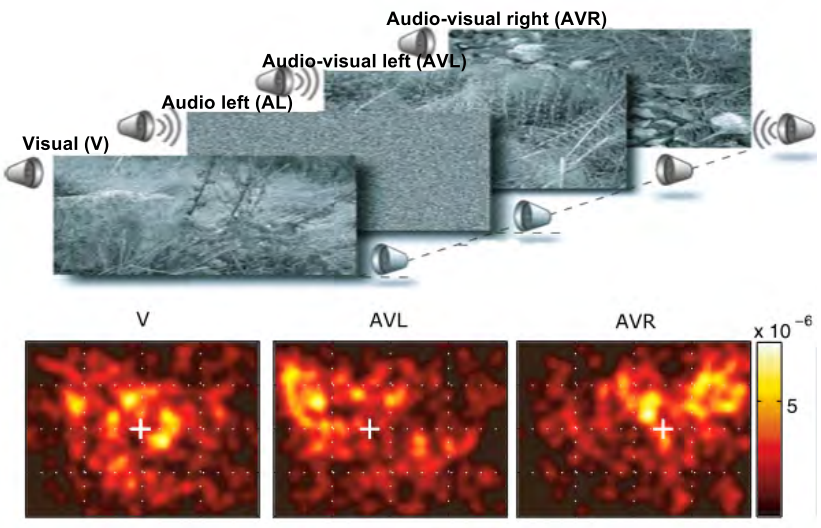


Figure 1.13 – **Haut :** conditions expérimentales. Scène naturelle statique sans aucun son (V), avec un son localisé en haut à gauche (AVL) ou à droite (AVR). **Bas :** cartes de densité des positions oculaires enregistrées dans les conditions expérimentales correspondantes. Les croix blanches représentent les centres de gravité des distributions le long de l'axe horizontal. Adapté de [Onat *et al.* 2007].

dernière. Une récente étude [Staufenbiel *et al.* 2011] a généralisé ce résultat aux stimuli dynamiques : les observateurs repéraient plus facilement le changement de trajectoire d'un point plongé dans un nuage d'autres points au mouvement uniforme lorsque ce changement était accompagné d'un bref son.

Il existe de nombreuses autres jolies preuves de l'intégration audiovisuelle, comme l'effet d'un simple bruit blanc sur la durée perçue d'un stimulus visuel [Gebhard & Mowbray 1959, Shipley 1964, Walker & Scott 1981, Welch *et al.* 1986, Recanzone 2003], ou encore sur la trajectoire perçue d'un objet visuel [Sekuler *et al.* 1997, Watanabe & Shimojo 1998, Meyer & Wuerger 2001, Freeman & Driver 2008, Väljamäe & Soto-Faraco 2008, Dufour *et al.* 2008, Hidaka *et al.* 2009]. Voir également la perception audiovisuelle de la parole et ses fameuses illusions telles que la ventriloquie ou l'effet McGurk, chapitre 4.

Ces résultats indiquent que l'information spatiale n'est pas nécessaire à l'intégration audiovisuelle, et que la synchronie temporelle des deux stimuli peut suffire. Cependant, tous ces résultats ont été obtenus à partir de stimuli saillants présentés dans un milieu confiné et artificiel. Se généralisent-ils en dehors de ces conditions ?

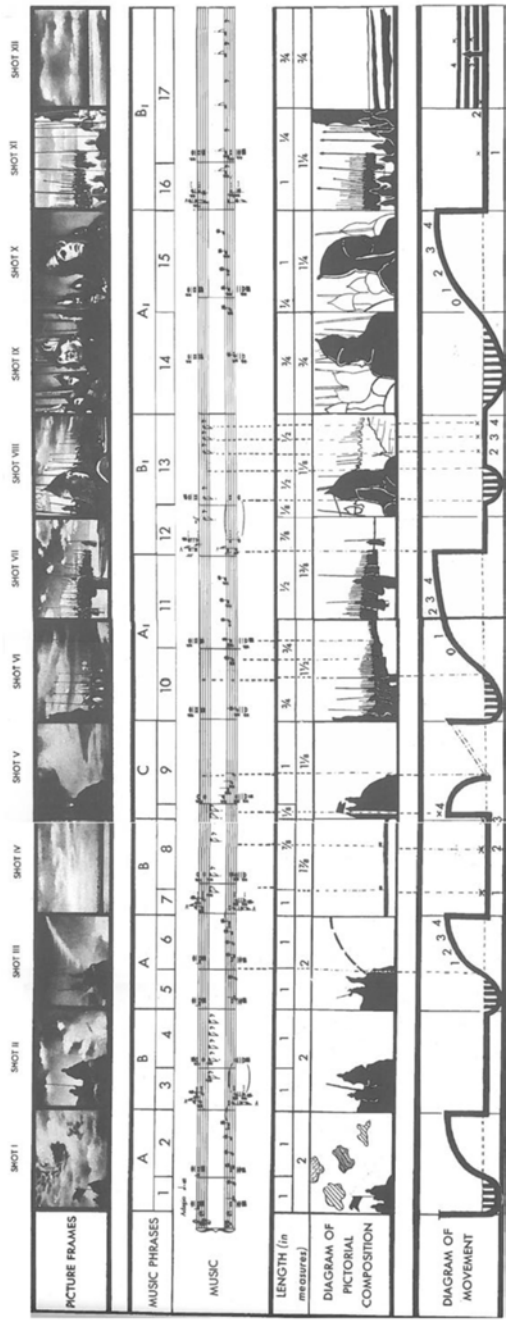


Figure 1.14 – Diagramme des correspondances audiovisuelles dans *Alexandre Nevsky* (Eisenstein, 1938). De haut en bas : frame représentative de chaque plan, phrase musicale, partition, durée (en mesures), composition picturale, et scanpath prédit par le réalisateur. Extrait de [Eisenstein 1943] (p. 148), repris dans [Smith 2014].

1.4.3 Scènes complexes

Dans la section précédente, nous avons vu de nombreuses preuves de l'intégration audiovisuelle utilisant des stimuli artificiels. Ici, nous faisons état des rares études mettant quantitativement en évidence l'intégration audiovisuelle dans des scènes complexes.

Une équipe de chercheurs allemands a enregistré les mouvements oculaires de 42 personnes visionnant des images statiques représentant des scènes naturelles (paysages) associées à des sons latéralisés (haut, bas, droite, gauche) simples, ou sans aucun son [Onat *et al.* 2007, Quigley *et al.* 2008]. Les auteurs ont constaté une déviation du regard en direction de la source sonore, comme le montre la Figure 1.13. De plus, ce biais était d'autant plus fort que la saillance visuelle à l'emplacement du son était importante, ce qui signifie que la densité de positions oculaires dépend à la fois de la saillance visuelle et de la saillance sonore. Au moyen de régressions linéaires multiples, les auteurs montrent qu'une combinaison linéaire des cartes de saillance unimodales constitue une bonne approximation de leurs données.

Cependant, ces résultats, s'ils présentent une avancée indiscutable vers une étude plus écologique de l'intégration audiovisuelle, restent difficilement généralisables. Dans la nature, le son est le plus souvent lié à un mouvement, à un déplacement, une agitation, bref à une scène dynamique. Rares sont les sons immobiles : sonnerie du téléphone, haut-parleur, grondement d'un lointain torrent, ils se comptent sur les doigts

de la main.

La dynamique temporelle du son et son association à l'image pour guider l'attention a été théorisée dès les débuts du cinéma. Le réalisateur russe Sergueï Eisenstein a proposé une analyse plan par plan d'une séquence de son film *Alexandre Nevsky*[8], incluant une prédiction de ce qu'il pensait être le scanpath idéal de son public. L'hypothèse d'Eisenstein était que le scanpath est déterminé par les correspondances audiovisuelles entre les changements dans la bande-son et la composition visuelle des plans (Figure 1.14).
Plus de 75 ans plus tard, Tim Smith, un chercheur londonien, a cherché à vérifier le bien fondé de cette hypothèse [Smith 2014]. Il a enregistré les mouvements oculaires de 26 participants regardant cette séquence, avec et sans la bande-son associée. Les résultats n'ont pas permis de confirmer l'intuition du réalisateur russe, aucune différence significative n'ayant été mesurée entre les deux conditions expérimentales. Par contre, le scanpath moyen correspondait partiellement avec celui imaginé par Eisenstein. Ceci laisse penser que ce sont avant tout les indices visuels qui permettent au réalisateur de guider le regard de son public. Mais il ne s'agit là que de l'influence de la musique qui, même si composée par Prokofiev, n'est pas forcément représentative des correspondances audiovisuelles dont nous sommes chaque jour témoins.
D'autres études oculométriques ont exploré l'influence du son lors de l'exploration de scènes naturelles dynamiques dans un contexte de perception audiovisuelle de la parole, nous en parlerons en détail au chapitre 4. Citons également le travail de thèse menée par Guanghan Song, qui a notamment établi une taxonomie des sons les plus susceptibles d'attirer l'attention lors de l'exploration de vidéos [Song 2013]. Mais nous reviendrons aussi sur ses travaux au cours de la discussion des chapitres 2 et 4.

1.4.4 Applications

Les modèles de prédiction de l'attention (visuels, auditifs ou audiovisuels) ont d'innombrables applications. Citons dans le désordre la segmentation d'objets (sonores ou visuels), la classification ou la mise en correspondance de scènes, la reconnaissance et la poursuite d'objets, la détection de changement de plan, la vidéo-surveillance, les prothèses rétiniennes, l'analyse d'images médicales, le marketing, l'interaction homme-machine... Pour une liste quasi-exhaustive accompagnée des références adéquates, se reporter à [Borji & Itti 2013]. Dans cette section, nous décrivons deux applications ayant mené au développement de modèle de saillance audiovisuelle *adhoc*.

8. Sergueï Eisenstein et Dmitri Vasilyev, *Alexandre Nevsky* (1938).

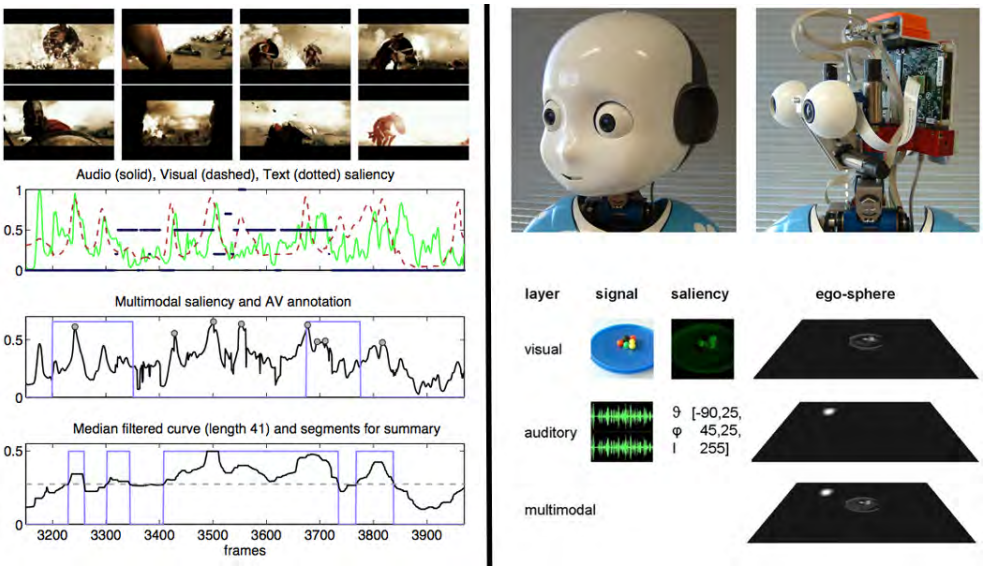


Figure 1.15 – Gauche : Résumé automatique de vidéos. De haut en bas : frames représentatives de la vidéo traitée, courbes de saillance sonore, visuelle et textuelle, courbe de saillance multimodale. Les rectangles bleus repèrent les périodes au dessus d'un certain seuil. Extrait de [Evangelopoulos *et al.* 2013]. **Droite :** Orientation de robots humanoïdes. En haut, la tête du robot humanoïde iCub, avec 6 degrés de liberté. En bas, les cartes de saillance visuelle et sonore décidant de l'orientation du robot. Adapté de [Ruesch *et al.* 2008].

1.4.4.1 Résumé automatique de vidéos

100 heures de vidéo sont mises en ligne chaque minute sur YouTube [9]. La constitution de ces immenses bases de données crée un besoin urgent d'outils d'exploration et d'indexation automatiques performants. Les algorithmes de résumés de vidéo, extrayant automatiquement les quelques frames les plus saillantes, entrent dans cette catégorie. Une équipe de recherche grecque a proposé un algorithme calculant trois courbes de saillance à partir des informations visuelle (intensité, couleur, orientation), sonore (modulations temps-fréquence) et textuelle (sous-titres) [Rapantzikos & Evangelopoulos 2007, Zlatintsi *et al.* 2012, Evangelopoulos *et al.* 2013]. Ces trois courbes sont ensuite combinées linéairement en une courbe de saillance multimodale. Les frames dont la saillance se situe au dessus d'un certain seuil sont choisies pour constituer le résumé de la vidéo. Ces différentes étapes sont résumées sur la partie gauche de la Figure 1.15.

1.4.4.2 Orientation spatiale de robots humanoïdes

Que ce soit à des fins industrielles ou récréatives, les robots humanoïdes sont de plus en plus présents dans notre environnement [Fong *et al.* 2003]. Afin d'amé-

9. `https://www.youtube.com/yt/press/fr/statistics.html`

liorer leurs interactions avec l'environnement, il est indispensable de les doter d'un "système perceptif" performant. Dans ce but, plusieurs équipes ont adopté une démarche biologiquement plausible basée sur l'estimation de saillance multimodale. Dans [Ruesch *et al.* 2008], une carte de saillance visuelle est calculée à partir de l'information d'intensité, de couleur et de mouvement. Puis une carte de saillance sonore à deux dimensions spatiales est calculée à partir des différences interaurales de phase et de niveau, évoquées section 1.3.2. Ces deux cartes sont ensuite combinées (opérateur *max*) en une carte de saillance multimodale guidant les yeux et le cou du robot. Par la suite, d'autres attributs plus haut niveau ont pu être rajoutés afin de rendre le robot plus "social" : reconnaissance de visage, d'expressions corporelle et faciale, de parole... [Zaraki *et al.* 2014].

1.5 Positionnement du problème

Au vu des résultats présentés dans ce chapitre, il semble qu'à l'heure actuelle, nous en sachions beaucoup sur les mécanismes attentionnels de la modalité visuelle, un peu sur ceux de la modalité sonore, mais nous ne comprenons pas vraiment comment les deux coopèrent lors de l'exploration de scènes naturelles. Dans ce manuscrit, nous abordons le problème en faisant trois choix théoriques et expérimentaux fondamentaux.

1. Tout d'abord, nous utilisons l'oculométrie pour quantifier l'effet du son sur l'exploration visuelle. Souple d'utilisation, cette méthode a en commun avec la vision une excellente résolution spatiale (0.01°d'angle visuel), et avec l'audition une excellente résolution temporelle (1000 Hz). Ceci en fait un outil particulièrement adapté à l'étude des mécanismes de l'attention audiovisuelle.
2. Ensuite, nous nous intéressons avant tout aux mécanismes ascendants. Il semble en effet raisonnable de commencer par étudier les aspects les plus reproductibles et généralisables du problème, avant de s'intéresser à leur modulation par des facteurs de plus haut niveau.
3. Enfin nous utilisons l'information temporelle et physique des signaux sonores, laissant de côté leur information spatiale. Aussi, nous n'utilisons que des stimuli monophoniques. Nous avons fait ce choix pour deux raisons. D'abord, comme nous l'avons vu section 1.4.2.2, de nombreuses expériences ont montré que la spatialisation sonore ne semble pas indispensable à l'intégration audiovisuelle. Ensuite, assurer la spatialisation sonore dans un contexte expérimental est complexe, tant techniquement que financièrement.

Après ce premier chapitre dressant l'état de l'art, le manuscrit est organisé de la manière suivante, chaque "•" correspondant à un nouveau chapitre.

• Une des premières choses à faire est de vérifier que le son a effectivement une influence sur l'exploration visuelle de scènes naturelles dynamiques et, le

cas échéant, d'en comprendre les principaux ressorts. Grâce à une première expérience oculométrique, nous avons mesuré et comparé les mouvements oculaires de participants visionnant des vidéos au contenu visuel très varié, avec et sans bande-son. Nous présentons également un modèle de saillance sonore dont nous nous servons pour étudier les mouvements oculaires effectués au voisinage des principaux événements auditifs.

• Les résultats de cette première expérience, s'ils mettent effectivement en évidence une influence de la modalité sonore, ne permettent pas de quantifier précisément les interactions à l'œuvre entre les différents attributs visuels et sonores. Lors d'une deuxième expérience oculométrique, nous comparons plus précisément les stratégies d'exploration visuelle en fonction du type d'association entre contenus visuels et sonores. Nous nous servons de méthodes de modélisation statistique afin d'estimer l'importance de différents attributs visuels pour expliquer les données enregistrées.

• Cette deuxième expérience montre que les scènes contenant des visages sont celles où l'influence du son est la plus forte. Nous nous penchons plus spécifiquement dessus dans le quatrième chapitre. Nous commençons par brosser un état de l'art de la perception audiovisuelle de la parole. Ces nombreux travaux nous permettent d'approfondir l'analyse et l'interprétation des explorations des scènes de conversation.

• Le dernier chapitre approfondi la compréhension des processus gouvernant l'exploration de scènes dynamiques sociales. Il présente une troisième expérience oculométrique utilisant de nouvelles scènes de conversations. Il s'agit ici de scènes de réunions de travail, mieux contrôlées que les précédentes (lieux fermés), et rendant plus précise la quantification des attributs audiovisuels des différents protagonistes. Ces résultats nous permettent de proposer un modèle de saillance audiovisuelle approprié à ce type de scène.

Tous les stimuli et mouvements oculaires issus de ces trois expériences sont disponibles à l'adresse suivante : `http://www.gipsa-lab.fr/~antoine.coutrot/`

Chapitre 2

Influence globale du son sur l'exploration visuelle

Sommaire

Dans ce chapitre, nous nous intéressons aux effets que la présence ou l'absence d'information sonore pourrait induire sur l'exploration libre de scènes naturelles dynamiques. Comme nous l'avons vu au chapitre précédent, cette question a priori simple n'a pratiquement jamais été abordée avec des scènes naturelles. La quasi totalité des expériences oculométriques utilisant des vidéos n'a jamais considéré leurs bandes-son et présentait aux participants des films muets, ce qui est loin d'être une condition écologique.
Nous présentons l'expérience 1, durant laquelle les mouvements oculaires de 40 participants ont été enregistrés alors qu'ils regardaient des vidéos au contenu très varié avec leurs bandes-son associées, ou sans aucun son. Dans un premier temps nous décrivons le protocole mis en place, les métriques utilisées pour comparer les mouvements oculaires enregistrés, et les résultats obtenus en moyenne ainsi qu'en fonction du temps. Dans un second temps, nous présentons deux modèles de saillance sonore dont le but est de repérer les événements auditifs les plus saillants. Ces modèles nous permettent de vérifier si les caractéristiques des mouvements oculaires diffèrent à proximité des événements sonores caractérisés.

2.1 Expérience 1

2.1.1 Hypothèses

Nous formulons l'hypothèse que supprimer l'information sonore contenue dans la bande-son d'une vidéo modifie son exploration visuelle. Par information sonore, nous entendons le contenu haut niveau (sémantique) et bas niveau (propriétés physiques du signal sonore). Nous ne prenons pas en compte la spatialisation du son, et travaillons donc avec des bandes-son monophoniques. Nous supposons que la perte de l'information sonore va entraîner une plus grande variabilité entre les positions oculaires des différents observateurs, ces derniers étant moins guidés dans leur exploration.

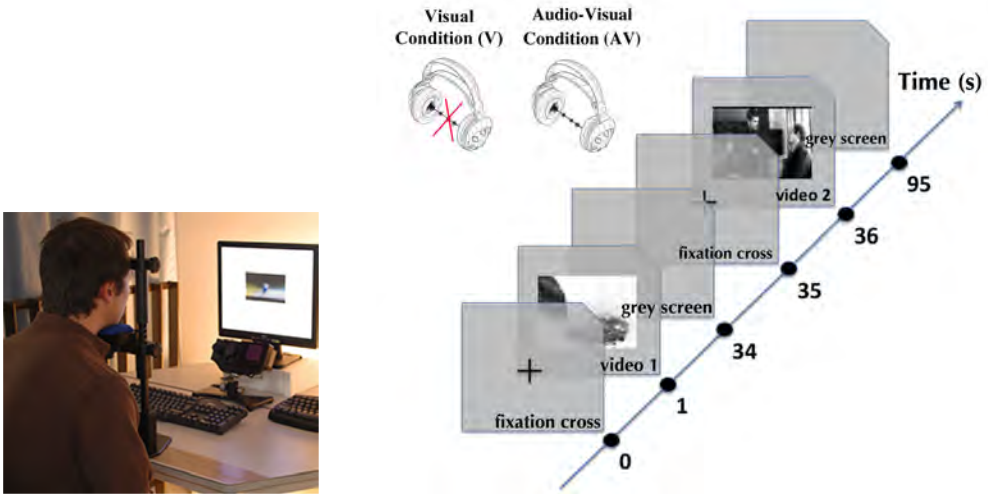
2.1.2 Design Expérimental

2.1.2.1 Participants

40 étudiants de l'université de Grenoble-Alpes ont participé à l'expérience : 26 hommes et 14 femmes, âgés entre 20 et 29 ans ($M = 25.3\,; SD = 2.7$). Les participants étaient naïfs quant au but de l'expérience, et avaient pour consigne de regarder librement et attentivement les vidéos présentées. Tous les participants étaient de langue maternelle française et avaient une ouïe normale. Leur vue était normale ou corrigée à la normale. Chacun a donné son consentement éclairé à prendre part à l'expérience.

2.1.2.2 Dispositif

Les participants étaient assis à 57 cm d'un écran à tube cathodique (CRT) de 21 pouces ViewSonic G220f, avec une résolution de 1024×768 pixels et un taux de rafraîchissement de 75 Hz. Leurs têtes étaient confortablement stabilisées au niveau du menton par un petit coussinet, et du front par un appui-tête (voir Figure 2.1a). Le son était diffusé au moyen d'un casque dynamique fermé à couplage circumaural (qui englobe l'oreille) Sennheiser HD280 Pro, 64Ω. Ce casque était porté tout au long de l'expérience, même en l'absence de stimulus sonore, ses 32 dB d'atténuation assurant une bonne isolation de potentiels bruits environnants distracteurs. Les mouvements oculaires étaient enregistrés par un oculomètre Eyelink 1000 (SR Research), avec une fréquence d'échantillonnage de 1000 Hz et une résolution spatiale nominale de 0.01 degré d'angle visuel. Nous n'enregistrions que l'œil directeur. Ainsi, une position oculaire était enregistrée toutes les millisecondes, en mode de suivi "pupille et reflet cornéen".



(a) (b)

Figure 2.1 – **(a)** Dispositif expérimental. La caméra de l'oculomètre est placée sous l'écran, face au participant. **(b)** Décours temporel d'un essai de l'expérience 1. La présentation d'une croix de fixation (1s), d'une vidéo, puis d'un écran gris (1s) est répétée pour les 50 vidéos. Un bloc de 25 vidéos est présenté sans son (condition Visuelle), et un autre bloc de 25 vidéos est présenté avec les bandes-son originales (condition AudioVisuelle).

2.1.2.3 Stimuli

Nous avons choisi 50 vidéos avec leurs bandes-son originales. Les vidéos sont extraites de films réalisés par des professionnels (films d'action, drames, documentaires). Elles ont une résolution de 720×576 pixels (30×24 degrés d'angle visuel), une fréquence de 25 images par seconde, et durent entre 7.8 s et 65.3 s ($M = 27.7; SD = 12.9$). Une vidéo peut contenir plusieurs plans. Au total, 163 plans ont été recensés dans l'ensemble des vidéos ($M = 8.7; SD = 7.2$). Les bandes-son sont monophoniques et échantillonnées à 48000 Hz. Si à l'origine, le son était stéréophonique, nous avons ajouté les deux canaux et envoyé la somme dans chaque écouteur. Lorsque la bande-son contient de la parole, c'est toujours en français. Une illustration de chacun des stimuli utilisés est disponible en Annexe A.

2.1.2.4 Protocole

L'expérience a été créée grâce au logiciel SoftEye, développé au laboratoire [Ionescu *et al.* 2009]. Ce programme permet de synchroniser la présentation des stimuli avec l'enregistrement des mouvements oculaires. Lors d'une expérience, il inscrit chronologiquement dans un même fichier *eyedatafile* tous les événements oculaires (saccades, fixations et clignements (*blinks*)). Un essai se déroulait de la manière suivante. Un fond gris était présenté pendant une seconde, puis une croix de fixation

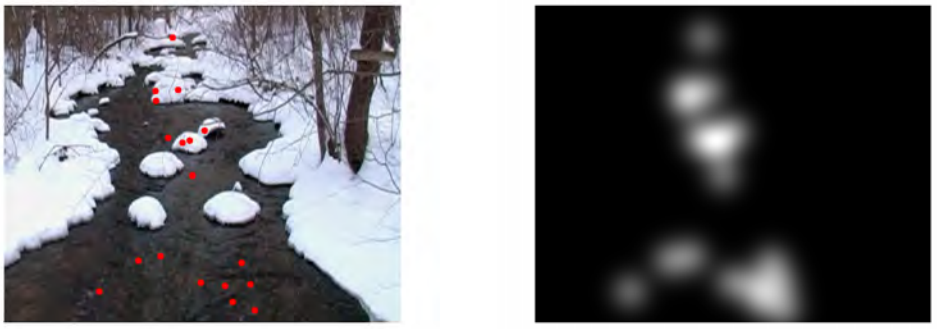
centrale apparaissait. Si le regard du participant était bien centré, la croix disparaissait et une vidéo était jouée, sur le même fond gris. Cette séquence était répétée pour les 50 vidéos, comme l'illustre la Figure 2.1b. Chaque expérience était précédée par une procédure de calibration, durant laquelle les participants devaient stabiliser leur regard sur 9 cibles réparties sur une grille 3×3 occupant tout l'écran. Une correction de la dérive du regard était effectuée entre chaque vidéo, et une nouvelle calibration était effectuée au milieu de l'expérience ou si la dérive excédait 0.5 degré. Afin d'éviter tout effet d'ordre ou de fatigue attentionnelle, les vidéos étaient présentées dans un ordre aléatoire. Les 20 premiers participants ont vu la première moitié des vidéos dans la condition AudioVisuelle (avec les bandes-son originales), et la seconde moitié dans la condition Visuelle (sans aucun son). L'ordre des conditions a été contrebalancé pour les 20 participants suivants. Une pause était systématiquement proposée au milieu de l'expérience, et les participants étaient informés qu'ils pouvaient se reposer entre chaque vidéo. Le cas échéant, une calibration était à nouveau effectuée à la reprise de l'expérience. Une expérience durait environ une demi-heure. Au final, chaque vidéo a été vue par 20 participants dans la condition Visuelle, et par 20 autres participants dans la condition AudioVisuelle.

2.1.2.5 Organisation des données

Nous n'avons analysé que les données issues d'un des deux yeux de chaque participant (de préférence l'œil directeur).

Positions oculaires Comme l'oculomètre enregistre une position oculaire toutes les millisecondes et qu'une frame dure 40 ms (25 frames par seconde (fps)), 40 positions oculaires par frame et par participant sont enregistrées. Par la suite, une "position oculaire" désignera la position médiane des 40 positions brutes. Il y aura donc une position oculaire par frame et par participant, sauf pour les frames correspondants aux blinks des participants.

Saccades, Fixations et Blinks Une saccade est détectée par le logiciel de l'oculomètre au moyen de trois seuils différents : un seuil de vitesse (30 degrés/s), d'accélération (8000 degrés/s^2) et de déplacement saccadique (0.15 degré). Une fixation est détectée dès lors que la pupille est visible et qu'aucune saccade n'est en cours. Les *blinks*, quant à eux, sont détectés comme des saccades avec une occlusion totale ou partielle de la pupille.
Ces mouvements oculaires ont été séparés en deux jeux de données : d'une part ceux enregistrés dans la condition AudioVisuelle (avec les bandes-son originales), et d'autre part ceux enregistrées dans la condition Visuelle (sans aucun son). Dans la section suivante sont présentées les différentes métriques utilisées pour comparer ces deux jeux de données.



(a) (b)

FIGURE 2.2 – **(a)** Les points rouges représentent les positions oculaires médianes de 18 participants. **(b)** Carte de densité des positions oculaires correspondantes.

2.1.3 Métriques

Dans cette section, nous présentons les métriques que nous allons utiliser pour caractériser les mouvements oculaires enregistrés lors de l'expérience 1. Nous les réutilisons également dans la suite du manuscrit, pour les expériences 2 et 3.

2.1.3.1 Dispersion

Pour estimer la variabilité des positions oculaires entre les différents participants, nous avons utilisé la "dispersion". Pour une frame regardée par n participants (donc pour n positions oculaires $\mathbf{p} = (x_i, y_i)_{i\in[1..n]}$), la dispersion D est définie de la manière suivante :

$$D(\mathbf{p}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{2.1}$$

En d'autres termes, la dispersion est la moyenne des distances euclidiennes entre les positions oculaires des n participants sur la frame considérée. Si tous les participants regardent au même endroit au même moment, la dispersion sera petite. Si au contraire les positions oculaires sont éparpillées sur la frame, la dispersion sera grande. Pour une frame donnée, deux types de dispersion ont été définis. D'une part, la dispersion "Inter", prenant en compte l'ensemble des positions oculaires enregistrées dans les deux conditions expérimentales. D'autre part, les dispersions "Intra", prenant en compte les positions oculaires dans chaque condition séparément.

2.1.3.2 Divergence de Kullback-Leibler

Comme nous l'avons vu au chapitre précédent (section 1.2.3.3), la divergence de Kullback-Leibler (DKL) peut être utilisée pour évaluer un modèle de saillance. Mais elle a aussi montré son utilité pour comparer des distributions de positions oculaires [Tatler *et al.* 2005, Le Meur *et al.* 2007, Quigley *et al.* 2008, Ho-Phuoc *et al.* 2012, Song *et al.* 2013, Le Meur & Baccino 2013]. Pour la situation qui nous intéresse, nous comparons les cartes de positions oculaires dans les deux conditions expérimentales (Visuelle et AudioVisuelle) pour une frame donnée. Ces deux cartes, Q^v et Q^{av} sont définies par l'équation 1.2. Elles ont donc la même taille qu'une frame (ici $p = 720 \times 576$ pixels), et sont normalisées de manière à constituer deux densités de probabilité (Figure 2.2). Leur DKL est donnée par

$$DKL(Q^V, Q^{AV}) = \frac{1}{2}\left(\sum_{i=1}^{p} Q_i^V \log \frac{Q_i^V}{Q_i^{AV}} + \sum_{i=1}^{p} Q_i^{AV} \log \frac{Q_i^{AV}}{Q_i^V}\right) \tag{2.2}$$

Pour une frame donnée, deux types de DKL ont été définis. D'une part, la DKL-inter, c'est-à-dire la divergence entre les cartes de positions oculaires enregistrées dans les deux conditions expérimentales : Q^v et Q^{av}. D'autre part, la DKL-intra, c'est-à-dire la divergence entre les positions oculaires au sein d'une même condition. La DKL-intra pourra être utilisée comme référence : si elle est inférieure à la DKL-inter, alors nous pourrons affirmer que les regards des participants ont été attirés par des régions différentes selon les conditions expérimentales. Pour calculer la DKL-intra, nous avons aléatoirement séparé les positions oculaires d'une condition expérimentale donnée en deux groupes de même taille, et avons calculé la DKL entre les cartes correspondant à ces derniers. Cette opération a été répétée 10 fois, et nous avons défini la DKL-intra comme la moyenne de ces 10 valeurs.

La dispersion et la DKL sont deux métriques complémentaires. La dispersion renseigne sur la variabilité entre les positions oculaires, mais ne dit rien sur les régions de l'image effectivement regardées par les participants. Pour la DKL, c'est le contraire.

2.1.3.3 Distance au centre

Comme son nom l'indique, la distance au centre mesure la distance euclidienne moyenne entre le centre du stimuli (x_c, y_c) et les positions oculaires des différents participants ($\mathbf{p} = (x_i, y_i)_{i \in [1..n]}$).

$$DC(\mathbf{p}) = \frac{1}{n}\sum_{i=1}^{n} \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2} \tag{2.3}$$

Cette métrique reflète le biais de centralité, c'est-à-dire la tendance que l'on a à regarder davantage au centre d'une image plutôt qu'à sa périphérie (voir section 1.2.2.3).

2.1.4 Résultats

Dans cette section, nous comparons les mouvements oculaires enregistrés dans la condition AudioVisuelle (AV, avec les bandes-son originales) et Visuelle (V, sans aucun son). Nous comparons les amplitudes de saccade et les durées de fixation effectuées dans les deux conditions expérimentales. Nous caractérisons la distribution des positions oculaires au moyen de la dispersion, de la distance au centre et de la divergence de Kullback-Leibler (DKL). D'une part, nous avons choisi de comparer la moyenne de ces métriques sur chacun des 163 plans de notre expérience, car comme il a été rappelé dans l'introduction 1.2.2.2, le plan est l'unité de base de la vidéo, et tout changement de plan influence grandement l'exploration visuelle. D'autre part, nous avons également analysé l'évolution temporelle des métriques au cours d'un plan. Les données de quatre sujets ont été exclues de l'analyse suite à des problèmes lors l'enregistrement (fichiers *eyedatafile* corrompus).

2.1.4.1 Analyse globale

Durées de fixation et amplitudes de saccade

Les durées de fixation et les amplitudes de saccade suivent une distribution asymétrique positive, comme l'illustre la Figure 2.3. Les amplitudes de saccade médianes ont été calculées pour chacun des 36 participants. Une ANOVA à un facteur intra indique qu'elles sont significativement supérieures dans la condition AV que dans la condition V ($F(1,35) = 5.6$, $p = .02$). Les durées de fixation médianes ont également été calculées pour chacun des 36 participants, et une ANOVA à un facteur intra ne trouve pas de différence entre les deux conditions expérimentales ($F(1,35) = 2.7$, $p = .11$).

Dispersion : la variabilité des positions oculaires

Pour chaque frame, nous calculons 3 valeurs de dispersion :
- Inter V-AV, à partir de l'ensemble des positions oculaires enregistrées, toutes conditions expérimentales confondues (soit 36 positions oculaires par frame),
- Intra AV, à partir des positions oculaires enregistrées dans la condition AudioVisuelle (soit 18 positions oculaires par frame),
- Intra V, à partir des positions oculaires enregistrées dans la condition Visuelle (soit 18 positions oculaires par frame).

Ces dispersions sont ensuite moyennées sur l'ensemble des frames de chacun des 163 plans, et nous comparons les moyennes de ces 163 valeurs (voir Figure 2.4a).

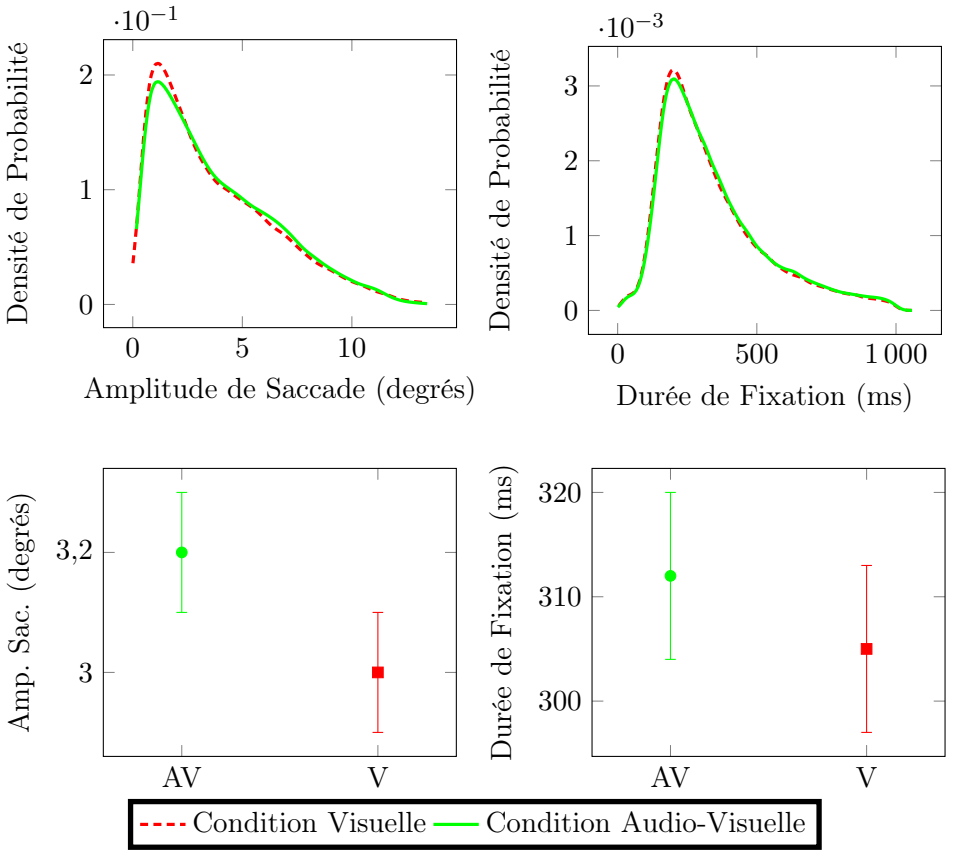


Figure 2.3 – **Haut :** Distributions des amplitudes de saccade et des durées de fixation dans les deux conditions expérimentales. Les densités de probabilité ont été estimées sur 100 points régulièrement espacés couvrant l'ensemble des données (fonction Matlab ksdensity). **Bas :** Amplitude de saccade et durée de fixation médianes, moyennées sur l'ensemble des participants. Les barres d'erreur correspondent aux erreurs standards.

Une ANOVA à un facteur intra (les conditions expérimentales : Inter V-AV, Intra AV, Intra V) révèle une différence significative entre les trois valeurs de dispersion [1] ($F(2,324) = 35.2$, $p < .001$), et des comparaisons a posteriori de Bonferroni indiquent que la dispersion Intra AV est inférieure aux dispersions Inter V-AV et Intra V ($p < .001$). Les dispersions Inter V-AV et Intra V ne sont pas significativement distinctes ($p = .88$). Ces résultats indiquent que les positions oculaires des différents participants sont en moyenne moins dispersées avec la bande-son originale que sans aucun son.

Distance au centre

Nous calculons pour chaque frame la distance au centre moyenne des positions

1. Une ANOVA avec les conditions expérimentales (V et AV) comme facteur intra et l'ordre de présentation (AV puis V ou V puis AV) comme facteur inter nous a permis de vérifier que l'ordre de présentation des conditions expérimentales n'a pas d'effet.

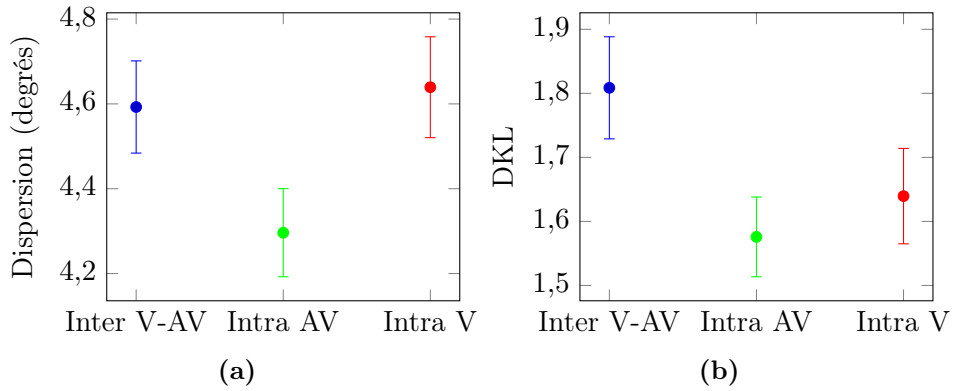


FIGURE 2.4 – (a) Valeurs moyennes de la dispersion en degrés angulaires : entre toutes les positions oculaires (Inter V-AV), entre les positions oculaires enregistrées dans la condition AudioVisuelle (Intra AV) et dans la condition Visuelle (Intra V). **(b)** Valeurs moyennes de la DKL entre les cartes de positions oculaires enregistrées dans la condition Visuelle et celles enregistrées dans la condition AudioVisuelle (Inter V-AV), au sein de la condition AudioVisuelle (Intra AV), et au sein de la condition Visuelle (Intra V). Les barres d'erreur correspondent aux erreurs standards.

oculaires enregistrées dans les conditions expérimentales V et AV, puis nous en prenons la moyenne sur chacun des 163 plans. Condition AV : M =2.3° ; SE = 0.1°. Condition V : M =2.2° ; SE = 0.1°. Une ANOVA à un facteur intra (les conditions expérimentales : Intra AV et Intra V) ne montre pas d'effet significatif des conditions expérimentales ($F(1,162) = 3.5$, $p = .06$).

Divergence de Kullback-Leibler

Pour chaque frame, nous calculons 3 valeurs de DKL :
- Inter V-AV, la DKL entre les cartes de positions oculaires enregistrées dans la condition Visuelle et celles enregistrées dans la condition AudioVisuelle (18 positions oculaires par carte),
- Intra AV, la DKL moyennée sur 10 tirages aléatoires de 2 groupes de positions oculaires enregistrées dans la condition AudioVisuelle (9 positions oculaires par carte),
- Intra V, la DKL moyennée sur 10 tirages aléatoires de 2 groupes de positions oculaires enregistrées dans la condition Visuelle (9 positions oculaires par carte).

Ces DKL sont ensuite moyennées sur l'ensemble des frames de chacun des 163 plans, et nous comparons les moyennes de ces 163 valeurs (voir Figure 2.4b). Une ANOVA à un facteur intra (les conditions expérimentales : Inter V-AV, Intra AV et Intra V) révèle une différence significative entre ces trois valeurs de DKL ($F(2,324) = 14.6$, $p < .001$), et des comparaisons a posteriori de Bonferroni indiquent que la DKL Inter V-AV est supérieure aux DKL Intra AV et Intra V ($p < .001$). Les DKL Intra AV et Intra V ne sont pas significativement distinctes ($p = 1$). Ces résultats indiquent que deux participants au sein d'une même condition expérimentale regardent davantage les mêmes régions que deux participants appartenant à deux conditions expérimentales différentes.

2.1.4.2 Analyse temporelle

Nous nous intéressons ici à l'évolution frame par frame des différentes métriques au cours d'un plan. La Figure 2.5, nous permet de constater que cette évolution temporelle a la même allure pour la dispersion, la distance au centre, et la DKL, et ce quelle que soit la condition expérimentale. Les premiers instants de l'exploration de nos 50 vidéos étant biaisés par la croix de fixation centrale, nous avons séparé les premiers plans (50) des autres (113). Cette évolution semble composée de 3 phases. Dans un premier temps (les 5 premières frames), les valeurs des 3 métriques restent stables (autour du minimum pour les premiers plans, autour du maximum pour les autres). Puis, entre les frames 5 et 25, ces métriques évoluent rapidement. Pour les premiers plans, elles croissent linéairement, pour les autres, elles décroissent brutalement jusqu'à leur minimum autour de la frame 10, puis croissent jusqu'à la frame 25. Enfin, à partir de la frame 25, ces métriques se stabilisent autour d'un plateau, jusqu'au changement de plan suivant. Pour chacun des 163 plans, nous avons pris la valeur moyenne de ces métriques sur les trois périodes temporelles définies ci-dessus : des frames 1 à 5, 5 à 25, et 25 jusqu'à la fin du plan. Une ANOVA à deux facteurs intra (les conditions expérimentales et les périodes temporelles) a été menée pour chacune des métriques. Il est à noter que les erreurs standards des DKL Intra V et Intra AV sont plus petites que celles de la DKL Inter V-AV car ces premières sont issues de la moyenne de 10 tirages aléatoires de groupes de positions oculaires de chaque condition expérimentale.

Pour la **dispersion**, il y a un effet principal des conditions expérimentales ($F(1,162) = 12.4$, $p < .001$) : la dispersion Intra V est supérieure à la dispersion Intra AV. Il y a également un effet principal des périodes temporelles ($F(2,324) = 30.3$, $p < .001$), et des comparaisons a posteriori de Bonferroni indiquent que la dispersion au cours de la troisième phase est supérieure à celle au cours des deux premières ($p < .001$). Il n'y a pas de différence significative entre les deux premières phases ($p = .29$).

Pour la **distance au centre**, il y a un effet principal des périodes temporelles ($F(2,324) = 17.5$, $p < .001$), la distance au centre est plus grande durant la seconde phase que durant la première ($p = .003$), et durant la troisième que durant la deuxième ($p = .03$). Par contre, il n'y a pas d'effet des conditions expérimentales ($F(1,162) = 1.15$, $p = .28$), ce qui est cohérent avec les résultats statistiques de l'analyse globale.

Pour la **DKL**, il y a un effet principal des conditions expérimentales ($F(2,324)=4.0$, $p = .02$), la DKL Inter V-AV est supérieure à la DKL Intra AV ($p = .02$). Les DKL Intra AV et Intra V ne sont pas significativement distinctes ($p = 1$), ni les DKL Inter V-AV et Intra V ($p = .12$). Il y a également un effet principal des périodes temporelles ($F(2,324) = 22.7$, $p < .001$), la troisième période est supérieure aux deux premières ($p < .001$), mais qu'il n'y a pas de différence significative entre les deux premières phases ($p = .055$).

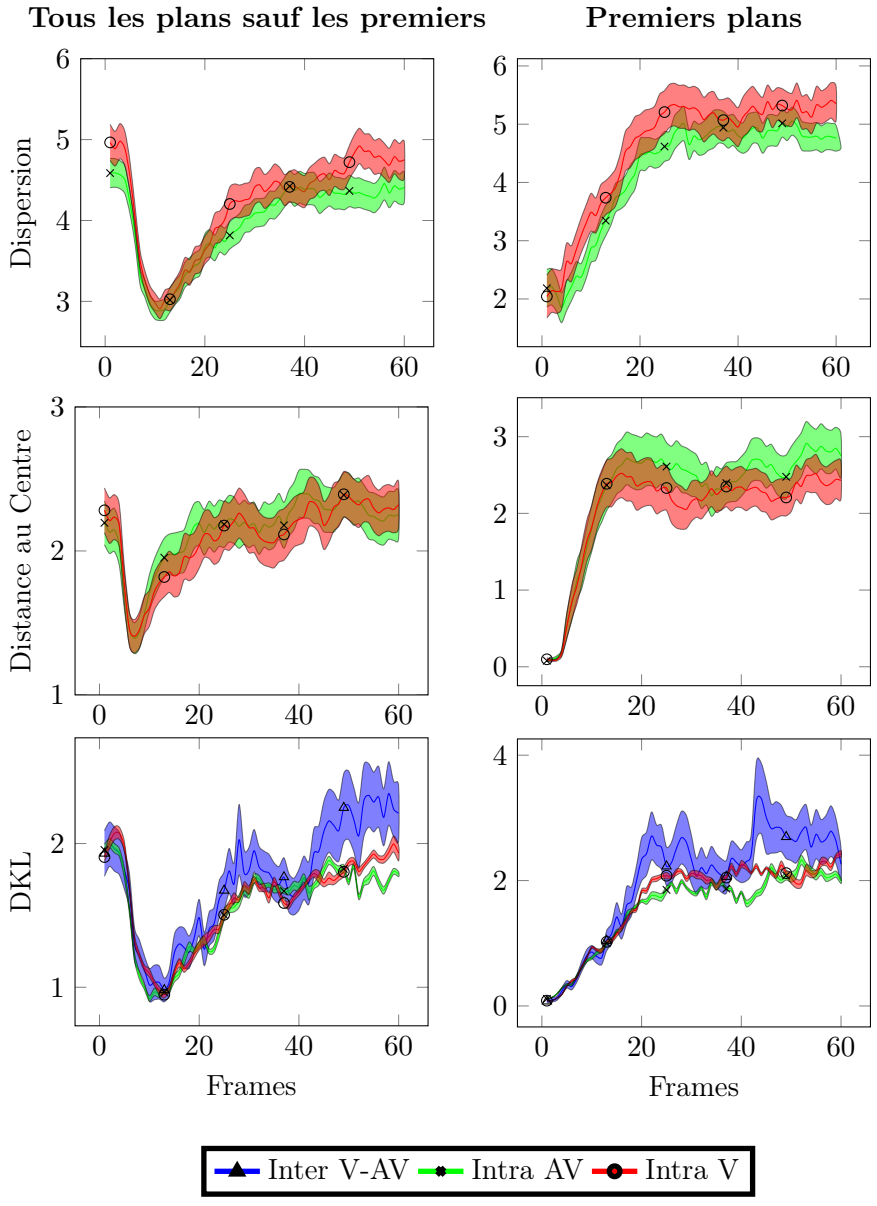


Figure 2.5 – Haut et milieu : évolution temporelle de la dispersion et de la distance au centre des positions oculaires dans les conditions AudioVisuelle (Intra AV) et Visuelle (Intra V). **Bas :** évolution temporelle de la DKL entre les cartes de positions oculaires de la condition Visuelle et celles de la condition AudioVisuelle (Inter V-AV), au sein de la condition AudioVisuelle (Intra AV), et au sein de la condition Visuelle (Intra V). Les premiers plans de chaque vidéo (50 plans, à droite) ont été séparés des autres (113 plans, à gauche). Les valeurs de la dispersion et de la distance au centre sont données en degrés angulaires, les barres d'erreur correspondent aux intervalles de confiance à 95%.

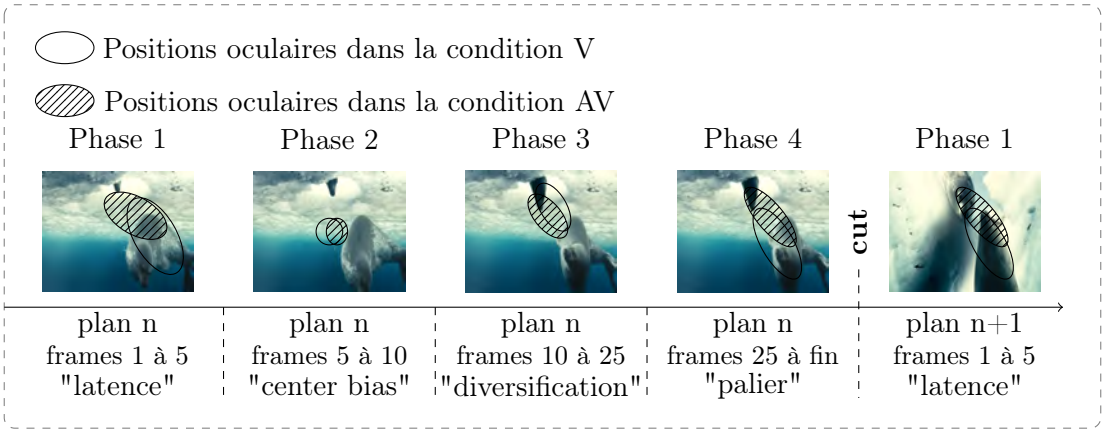


Figure 2.6 – Représentation schématique des quatre phases de l'exploration visuelle d'un plan dynamique. L'aire des ellipses est proportionnelle à la dispersion, et leur centre correspond à la position moyenne des positions oculaires dans les deux conditions expérimentales Visuelle (V) et AudioVisuelle (AV).

2.1.5 Discussion

Nous avons comparé les mouvements oculaires enregistrés lors de l'exploration libre de vidéos lorsque ces dernières sont accompagnées de leurs bandes-son originales (condition AV), ou d'aucun son (condition V). Nous avons mis en évidence une influence de la condition expérimentale sur les amplitudes de saccade et les positions oculaires au moyen de différentes métriques. Nous avons également montré que les valeurs de ces métriques varient considérablement au cours de l'exploration d'un plan. Aussi, avant de s'intéresser à l'influence du son proprement dite, il semble logique de revenir sur les différentes phases de l'exploration visuelle, indépendamment de la condition expérimentale.

Exploration visuelle de scènes dynamiques

Comme rappelé dans l'Etat de l'art (section 1.2.2.2), il a été montré que les scènes dynamiques suscitent chez leurs observateurs des mouvements oculaires d'une grande homogénéité, d'autant plus si les scènes en question ont été conçues et montées par des professionnels de l'image, comme c'est le cas pour nos stimuli [Goldstein *et al.* 2007, Hasson *et al.* 2008b, Dorr *et al.* 2010]. Ceci indique que le rythme, la dynamique d'une vidéo jouent un rôle important pour capter et guider le regard. En particulier, les changements de plans modifient l'exploration de manière assez radicale [Boccignone *et al.* 2005, Mital *et al.* 2010, Smith *et al.* 2012]. Ces transitions entre deux plans renouvellent, plus ou moins abruptement, l'ensemble de l'information présentée dans la scène, réinitialisant tous les processus et stratégies de traitement de l'information visuelle développés par les observateurs depuis le début de l'exploration [Wang *et al.* 2012]. Ces caractéristiques primordiales des

plans justifient que nous les ayons choisis comme unité temporelle de base pour mener nos analyses. Nous avons donc regardé l'évolution temporelle au cours d'un plan de trois métriques caractérisant chacune à leur façon les distributions de positions oculaires : la dispersion, qui caractérise la variabilité entre les positions des différents observateurs ; la distance au centre, qui caractérise la propension qu'ont les participants à s'éloigner du centre de l'écran pour explorer la scène ; et la divergence de Kullback-Leibler (DKL), qui caractérise la proximité des régions de la scène regardées par les observateurs. Dans la suite, nous ne discuterons que de l'évolution temporelle au cours des plans qui ne débutent pas une vidéo. En effet, les premiers instants d'exploration des premiers plans sont biaisés par la croix de fixation centrale qui précède chaque vidéo. Nous avons montré qu'après un changement de plan, l'évolution de la dispersion, de la distance au centre, et de la DKL ont la même allure. Nous décomposons cette dernière en quatre phases distinctes, dont nous proposons une interprétation (voir schéma 2.6).

Phase 1 : durant les 5 premières frames (200 ms), ces trois métriques restent stables, à leur maximum. Cette phase correspond au temps que mettent les participants pour réagir au brusque changement de plan, et commencer à explorer la nouvelle scène. Les métriques sont donc au niveau auquel elles étaient à la fin du plan précédant. Notons que ce délai est du même ordre de grandeur que la latence habituellement mesurée pour des saccades réflexes vers des cibles périphériques (120-200 ms) [Yang *et al.* 2002, Walker *et al.* 2006, Wu *et al.* 2010].

Phase 2 : entre les frames 5 et 10 (de 200 ms à 400 ms après le début du plan), les positions oculaires des participants convergent vers le centre de l'écran, induisant une baisse, naturellement de la distance au centre, mais également de la dispersion et de la DKL, les regards se regroupant autour d'une même zone. Ce comportement est lié au biais de centralité, phénomène bien identifié dans la littérature renvoyant notamment à l'idée que la position optimale pour débuter l'exploration d'une nouvelle scène est son centre (cf. 1.2.2.3) [Tatler 2007, Tseng *et al.* 2009].

Phase 3 : entre les frames 10 et 25 (de 400 ms à 1 s après le début du plan), les trois métriques augmentent linéairement. Cette phase correspond à la période précoce de l'exploration durant laquelle le regard des observateurs s'éloigne du centre de l'écran pour aller explorer les régions les plus saillantes de l'image. Rapidement, des stratégies d'exploration propres à chaque observateur se mettent en place et se diversifient, induisant une augmentation de la dispersion et de la DKL, tant Inter qu'Intra [Tatler *et al.* 2005].

Phase 4 : de la frame 25 jusqu'à la fin du plan (217 frames en moyenne), les trois métriques fluctuent autour d'une valeur moyenne, un palier. Ceci indique qu'environ une seconde après l'apparition du plan, les stratégies d'exploration cessent de se diversifier. Contrairement à l'exploration de scènes statiques, les scènes dynamiques sont constamment en mouvement, et de nouveaux objets saillants apparaissent régulièrement à l'image. De plus, il a été montré que les attributs dynamiques, comme le mouvement, attirent davantage le regard que les attributs statiques, comme le contraste [Carmi & Itti 2006, Mital *et al.* 2010, Smith & Mital 2013]. Ceci permet d'expliquer pourquoi ces métriques se stabilisent assez

rapidement : le constant renouvellement des informations présentes à l'écran limite le développement des processus descendants, les regards restant attirés par un nombre limité de régions saillantes [Wang *et al.* 2012].
Notons que ce qui précède est généralisable aux premiers plans, en considérant qu'au début de la première phase, les regards de tous les observateurs sont groupés au centre de l'écran (ou que le "plan précédant" est la croix de fixation centrale). Maintenant que nous avons une idée de la dynamique de l'exploration visuelle d'un plan d'une vidéo, nous pouvons nous intéresser à l'influence des conditions expérimentales sur cette dernière.

Influence du son

La plupart des paradigmes destinés à mettre en évidence l'effet du son sur les mouvements oculaires utilisent des stimuli artificiels, comme des cibles circulaires et des sons synthétiques. Ces études ont par exemple permis de démontrer que la perception d'un stimulus audiovisuel synchrone (temporellement et/ou spatialement) est plus précise et rapide que la perception d'un stimulus unimodal [Todd 1912, Corneil & Munoz 1996, Spence & Driver 1997, Corneil *et al.* 2002, Arndt & Colonius 2003]. Les rares études utilisant des scènes naturelles adoptent un point de vue spatial : elles cherchent à localiser la source sonore pour augmenter la saillance de la région correspondante [Onat *et al.* 2007, Quigley *et al.* 2008, Ruesch *et al.* 2008]. Dans cette expérience, nous avons poursuivi un but radicalement différent, puisque nous avons utilisé des bandes-son monophoniques, mettant de côté toute information spatiale. Nous avons fait l'hypothèse que l'information sonore contenue dans la bande-son, qu'elle soit haut niveau (sémantique) ou bas niveau (propriétés physiques du signal sonore) peut interagir avec l'information visuelle de la vidéo, par exemple en modifiant sa saillance, et donc le regard des observateurs. Nous observons que l'exploration visuelle, même si elle reste principalement déterminée par le contenu visuel (et notamment le montage des vidéos), est significativement influencée par la présence ou l'absence de son. Plus précisément, supprimer la bande-son a pour effet d'accroître la variabilité entre les positions oculaires des différents observateurs (dispersion), induisant des fixations dans des régions différentes de celles regardées si l'information sonore avait été présente (DKL). Une interprétation de ce phénomène est que le son renforce la saillance des objets visuels. Sans le son, les objets visuels sont moins saillants et attirent moins fortement le regard des observateurs, permettant une plus grande diversité des stratégies d'exploration visuelle. En somme, les observateurs sont moins bien "guidés" par des zones de forte saillance, ces dernières étant plus uniformément réparties. Ceci permet également d'expliquer les plus petites amplitudes de saccade enregistrées dans la condition Visuelle : sans le son, les observateurs feraient moins de saccades volontaires dirigées vers des régions précises de l'image. Cette interprétation est cohérente avec le résultat de certaines études utilisant des stimuli artificiels, comme le *pip and pop phenomenon* décrit section 1.4.2.2 [Van der Burg *et al.* 2008]. En bref, cette étude

montre que lors d'une tâche de recherche d'une cible au milieu de distracteurs, la présence d'un bref son "pip" monophonique synchronisé temporellement avec le changement de couleur de la cible améliore spectaculairement les performances des participants, comparé à la même tâche sans le son. Les auteurs interprètent ce phénomène comme un renforcement de la saillance de la cible visuelle par le signal sonore, ce qui est très proche de l'idée que nous venons de formuler pour expliquer nos résultats globaux.

Nous avons mis en évidence un effet global du son lors de l'exploration libre de scènes dynamiques. Nous allons à présent tenter de déterminer si cet effet est constant au cours du temps, ou s'il se renforce à proximité des événement sonores particulièrement saillants. Ceci nous renseignera sur la nature de l'interaction entre les informations visuelle et sonore. Est-elle ponctuelle, c'est-à-dire limitée aux relations immédiates entre sons et images à un instant T, ou s'étale-t-elle davantage dans le temps ?

2.2 Influence d'un événement sonore sur l'exploration visuelle

Dans cette section, nous décrivons en détail le modèle de saillance sonore que nous avons implémenté, pour extraire des bandes-son les événements sonores les plus saillants. Nous vérifions ensuite si les mouvements oculaires sont modifiés à proximité de ces derniers. Comme présenté dans l'état de l'art (1.3.3), le but d'un modèle de saillance sonore est de repérer les instants durant lesquels certains attributs d'un signal audio se distinguent de leur valeur moyenne, attirant ainsi l'attention des auditeurs [Bregman 1990]. Même si notre compréhension des mécanismes d'attention auditive reste limitée, certaines études ont établi que les modulations d'amplitude et de fréquence jouent un rôle majeur dans l'allocation attentionnelle [Kayser *et al.* 2005, Fritz *et al.* 2007]. Le Discrete Energy Separation Algorithm (DESA) est un modèle de saillance sonore basé sur les modulations temporelles d'amplitude et de fréquence du signal dans différentes bandes fréquentielles. La séparation du signal en plusieurs bandes de fréquence permet la capture de ces modulations même en présence de bruit, ce dernier étant souvent un facteur limitant lorsqu'il s'agit de scènes auditives complexes comme celles que nous utilisons [Bovik *et al.* 1993]. Cet algorithme a récemment été utilisé dans des études de détection de la parole en milieu bruité [Evangelopoulos & Maragos 2006], ou encore pour extraire automatiquement les frames les plus représentatives d'une vidéo à partir (notamment) de l'information sonore [Evangelopoulos *et al.* 2013].

2.2.1 Modèles de saillance sonore

2.2.1.1 le DESA

Afin de se doter d'un cadre temporel commun au traitement du son et de l'image, nous scindons le signal sonore en périodes de temps de même durée qu'une frame visuelle. Dans notre cas, nos bandes-son sont échantillonnées à 48 kHz et nos frames durent 40 ms : le signal audio est donc découpé en segment de L = 48000 x 0.04 = 1920 échantillons audio. En entrée du modèle, ces segments sont décomposés en N bandes de fréquence au moyen d'un banc de filtres de Gabor. Ces filtres ont souvent été utilisés pour modéliser les champs récepteurs des cellules de l'aire V1 du cortex visuel, car ils sont un bon compromis de résolution entre les domaines spatial et fréquentiel [Daugman 1980]. Temporellement, un filtre de Gabor correspond à un cosinus modulé par une gaussienne :

$$h_i(t) = exp(-\lambda_i^2 t^2) cos(\omega_i t) \tag{2.4}$$

avec $(\omega_i, \lambda_i)_{i\in[1..N]}$ respectivement les fréquences centrales et les largeurs de bande des filtres utilisés. Ces dernières ont été choisies de telle manière que deux filtres voisins s'intersectent à mi-hauteur :

$$\omega_i = \frac{3\Omega_c}{2^{i+1}} \text{ , et } \lambda_i = \frac{\omega_i}{2\sqrt{\ln 2}} \tag{2.5}$$

avec Ω_c la fréquence maximale du signal. Concrètement, les bandes-son, échantillonnées à 48 kHz, sont séparées en six bandes de fréquence centrées respectivement en $\omega_i \in \{281, 562, 1125, 2250, 4500, 9000\}$ Hz. Cette gamme spectrale couvre la plupart des sons audibles par l'oreille humaine (*e.g.* parole : entre 50 Hz et 8 kHz). Pour un échantillon audio k, l'énergie de Teager-Kaiser est définie de la manière suivante :
Teager-Kaiser energy :

$$\Psi[s(k)] = s^2(k) - s(k+1)s(k-1) \tag{2.6}$$

L'énergie de Teager-Kaiser est caractérisée par sa très bonne résolution temporelle, la rendant idéale pour l'analyse temporelle précise des signaux audio. Cette énergie est fréquemment utilisée pour détecter les modulations d'amplitude et de fréquence dans les signaux AM-FM [Teager 1980, Kaiser 1990].
Pour séparer le bruit environnant du signal d'intérêt, seule la bande de fréquence dont l'énergie de Teager-Kaiser est maximale est sélectionnée. Dans cette bande spectrale, l'énergie instantanée est décomposée en sa composante d'amplitude et de fréquence, selon les relations suivantes :
Amplitude instantanée :

$$|a[s(k)]| = 2\frac{\Psi(s(k))}{\sqrt{\Psi(\dot{s}(k))}} \tag{2.7}$$

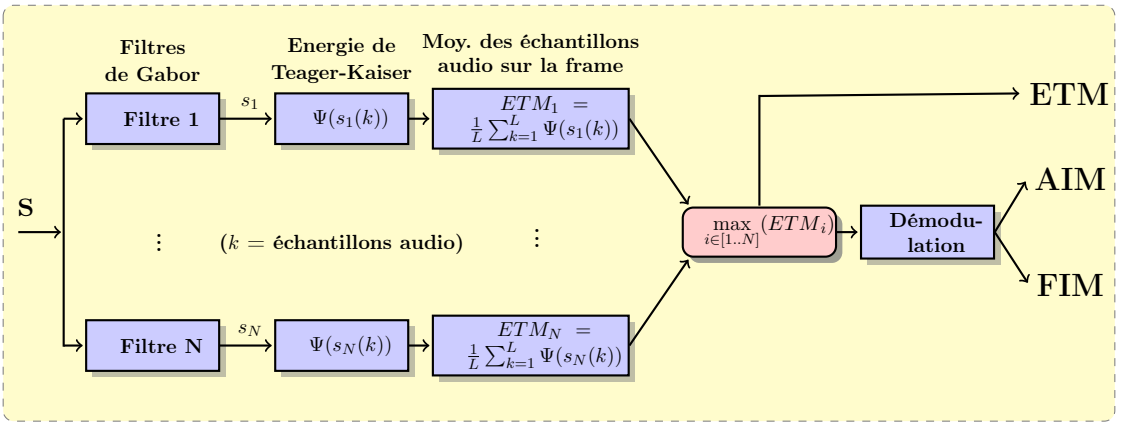


Figure 2.7 – Modèle de saillance sonore DESA. En entrée, le signal sonore s est séparé en N bandes de fréquence. L'énergie de Teager-Kaiser, donnée par l'équation 2.6, est moyennée sur les L échantillons audio k contenus dans la frame. La bande de fréquence avec l'ETM la plus grande est sélectionnée . Les amplitudes et fréquences instantanées moyennes (AIM et FIM) sont démodulées à partir de l'ETM grâce aux équations 2.7 and 2.8. La saillance sonore est une combinaison linéaire de ces trois attributs.

Fréquence instantanée :

$$f[s(k)] = \frac{1}{2\pi}\arcsin\left(\sqrt{\frac{\Psi[\dot{s}(k)]}{4\Psi[s(k)]}}\right) \tag{2.8}$$

avec $\dot{s}$ la dérivée du signal.
L'énergie de Teager-Kaiser, la fréquence instantanée et l'amplitude instantanée sont alors moyennées sur les L échantillons audio k contenus dans une frame, donnant l'Energie de Teager-Kaiser Moyenne (ETM), l'Amplitude Instantanée Moyenne (AIM) et la Fréquence Instantanée Moyenne (FIM). Ces trois attributs sont alors normalisés et combinés linéairement pour donner la valeur de saillance sonore de la frame vidéo correspondante (voir Figure 2.8).

$$S = \alpha_1\mathrm{ETM} + \alpha_2\mathrm{AIM} + \alpha_3\mathrm{FIM} \tag{2.9}$$

avec $\alpha_1 = \alpha_2 = \alpha_3 = \frac{1}{3}$. Notons qu'il est possible d'adapter la pondération de ces 3 attributs au type de scène sonore à traiter. Par exemple, s'il s'agit majoritairement de signaux présentant de grandes variations d'énergie (comme une dispute avec des éclats de voix), l'emphase sera davantage portée sur l'ETM. Si au contraire, la scène contient de grandes variations fréquentielles (comme l'effet Doppler de la sirène d'une ambulance), la FIM sera privilégiée. Ici, vue la grande variété des scènes sonores étudiées, nous avons choisi une pondération équilibrée, rendant le modèle aussi souple que possible.

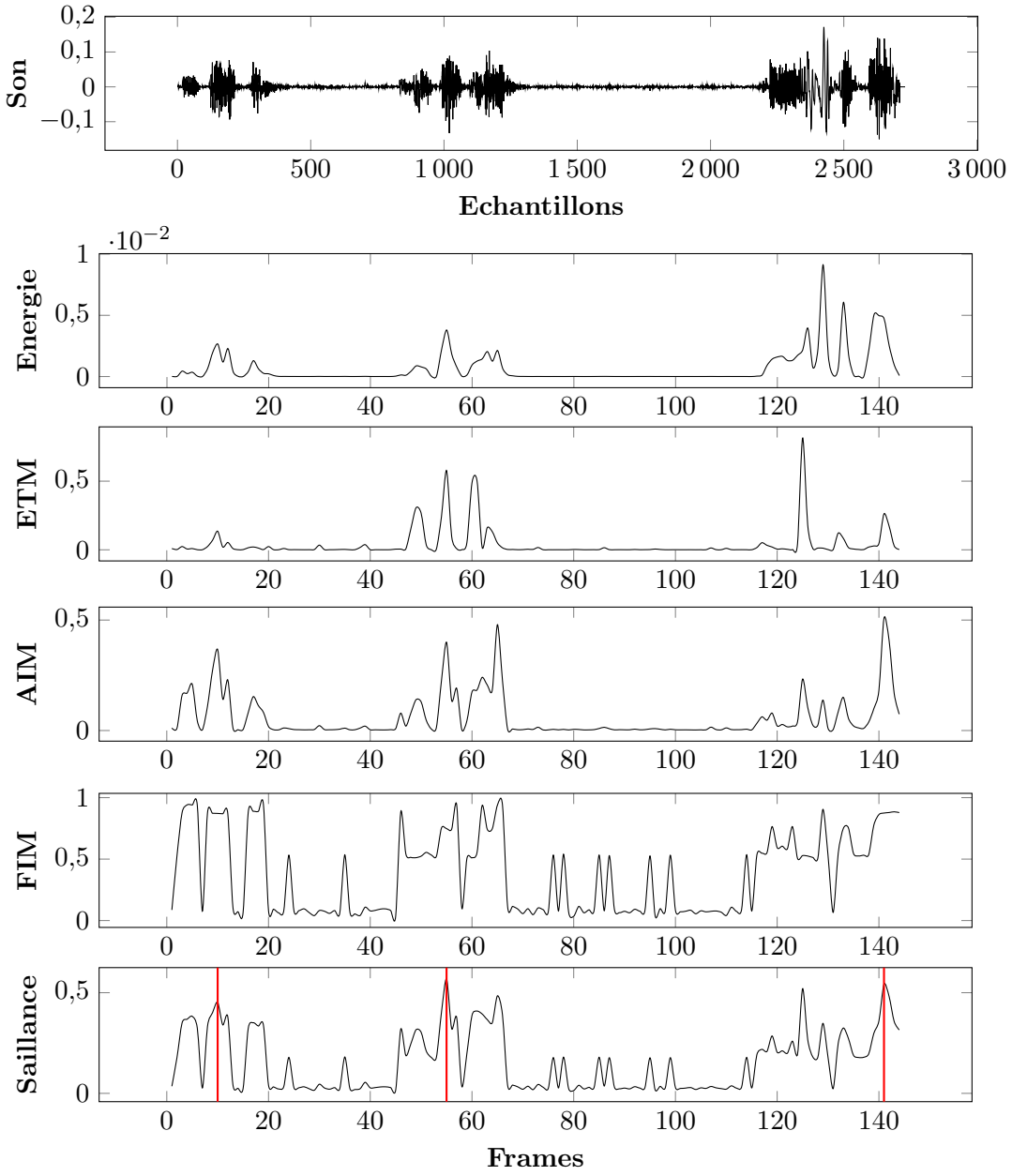


Figure 2.8 – Modèle de saillance sonore DESA. De haut en bas : (1) Signal d'entrée, séparé pour la suite en fenêtres de 40 ms (frames) ; (2) Energie du signal ; (3) Energie de Teager-Kaiser Moyenne ; (4) Amplitude Instantanée Moyenne ; (5) Fréquence Instantanée Moyenne ; (6) Saillance sonore issue de la moyenne de (3), (4) et (5). Les lignes verticales rouges du graphe inférieur repèrent les principaux pics de saillance.

2.2.1.2 Le modèle "Energie"

Nous comparons les résultats issus du DESA avec un modèle de saillance sonore bien plus simple, qui identifie la courbe de saillance à la courbe d'énergie du signal. Ainsi, la valeur de saillance S d'une frame comportant L échantillons audio devient :

$$S = \frac{1}{L}\sum_{k=1}^{L} s^2(k) \tag{2.10}$$

Une fois obtenue la courbe de saillance sonore, *via* le DESA ou le modèle Energie, il s'agit d'en extraire les principaux pics (les lignes verticales rouges en bas de la Figure 2.8). Nous avons normalisé le nombre de pic en fonction du temps : si un signal sonore dure N frames, les N/2 principaux pics sont conservés (0.5 par seconde). De plus, l'intervalle minimal entre deux pics est d'une seconde (25 frames). Ainsi, deux pics voisins sont tout de même suffisamment distants pour que le potentiel effet attentionnel de l'un n'affecte pas l'autre.

2.2.2 Evaluation qualitative des modèles

Section 1.3.4, nous avions présenté les différentes méthodes utilisées dans la littérature pour évaluer les modèles de saillance sonore. Ces méthodes souffraient toutes du même défaut : elles ne permettaient que d'évaluer la saillance d'un son isolé de son contexte général. Ici nous présentons une nouvelle méthode, qui bien que souffrant elle aussi de certaines limitations, permet de s'affranchir de ce problème.

2.2.2.1 Méthodologie

Stimuli - Trois jeux de pics étaient présentés : ceux issus du DESA, ceux issus du modèle Energie, et des pics aléatoires. Chaque jeu possédait le même nombre de pics, et chaque pic était distant d'au moins 25 frames de ses voisins. Nous avons sélectionné 14 bandes-son au contenu varié, présentant un total de 104 pics de saillance sonore ($M = 7.4$, $SE = 0.85$).
Participants et protocole - Nous avons demandé à 5 personnes d'écouter chacune des bandes-son (sans l'image), et de juger si les pics présentés correspondaient ou non à un événements particulièrement saillant. Les participants écoutaient les bandes-son à partir d'un ordinateur, *via* le même casque audio que celui utilisé lors de l'expérience 1. Pour chaque bande-son, trois courbes représentant les trois jeux de pics (DESA, Energie et aléatoires) étaient affichées, et un curseur vertical bleu indiquait la progression temporelle (voir Figure 2.9). Les participants devaient donc juger de manière binaire la saillance sonore de 104 pics x 3 modèles = 312 pics. S'ils le souhaitaient, les participants avaient la possibilité de mettre sur pause ou de rejouer un

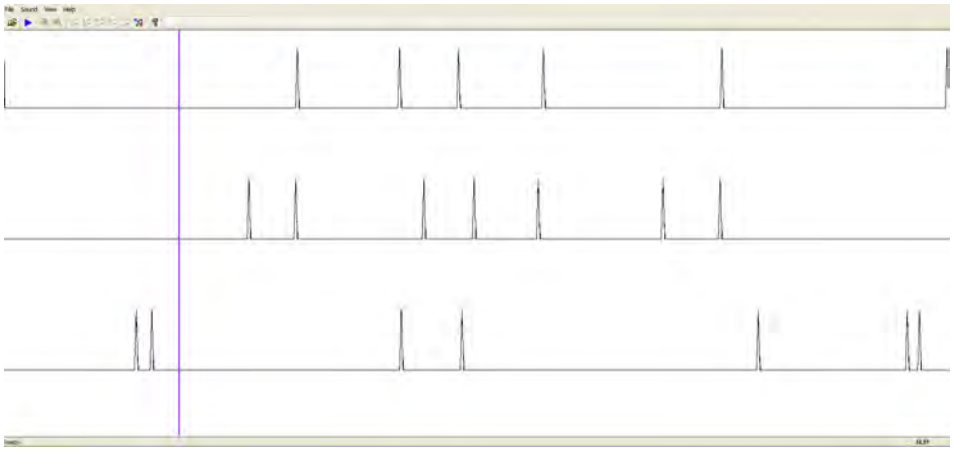


Figure 2.9 – Affichage utilisé pour évaluer l'efficacité des modèles de saillance sonore. Les trois lignes de pics noirs représentent les trois jeux de pics (DESA, Energie et aléatoire), et le curseur vertical bleu marque la progression temporelle de la bande-son.

Table 2.1 – Pourcentages, pour chaque jeu de pics (DESA, Energie et aléatoires), des événements jugés saillants par les participants ($M(\pm SE)$).

	Jeux de pics		
	DESA	Modèle Energie	Aléatoire
Efficacité (%)	43 (±4.9)	60 (±7.2)	20 (±4.1)

extrait. En moyenne, une évaluation complète prenait 45 minutes.

2.2.2.2 Résultats

Pour chaque bande-son, le pourcentage moyen de pics DESA, Energie et aléatoires jugés comme étant saillants par les participants est calculé (voir Table 2.1). Plus ce pourcentage est élevé, plus le jeu de pics de saillance sonore correspondant est proche de la réalité. Les pics de saillance issus du DESA et du modèle Energie sont jugés nettement meilleurs que les pics aléatoires (tests de Student respectifs : t(13)=5.5, $p < .001$ et t(13)=5.3, $p < .001$). De plus, les pics de saillance du modèle Energie sont jugés plus proches de la réalité que ceux du DESA (t(13)=-2.8, $p < .01$).

2.2.3 Résultats

Nous reprenons les résultats issus de l'analyse globale des mouvements oculaires enregistrés lors de l'expérience 1, dans la condition Visuelle et AudioVisuelle (section

2.1.4.1). Nous comparons ces mouvements oculaires "moyens" avec ceux effectués juste après deux types d'événements sonores particuliers :

1. les pics de saillance repérés par le DESA,
2. les pics de saillance repérés par le modèle Energie,

Pour chaque plan contenant au moins un pic DESA ou Energie (127 plans sur 163), nous comparons les valeurs de dispersion moyennées sur l'ensemble des frames du plan, avec celles moyennées sur une petite fenêtre temporelle suivant chaque événement sonore. Pour évaluer l'éventuelle persistance dans le temps des pics de saillance, nous comparons nos résultats pour des fenêtres de 5, 10, ou 25 frames. Nous avons également fait varier le nombre de pics considérés, en ne menant les analyses que sur les P principaux pics de la bande-son, avec P compris entre 3 et N/2, N étant le nombre de frames de la bande-son. Ci-dessous ne sont présentés que les résultats pour P=N/2, les résultats pour des valeurs de P inférieures étant similaires. A noter que plus P est petit, moins il y a de données, et plus la variance des résultats augmente.

Après les pics de saillance DESA Pour tester l'influence de la taille de la fenêtre d'analyse de la dispersion après les pics de saillance DESA, une ANOVA à deux facteurs intra (la condition expérimentale V ou AV ; la taille de la fenêtre temporelle 5, 10 ou 25 frames) a été menée. Elle révèle un effet principal de la condition expérimentale ($F(1,126)=31.8$, $p < .001$), conformément aux résultats de la section 2.1.4.1. Par contre, elle ne met en évidence aucun effet de la taille de la fenêtre temporelle ($F(2,124)=0.9$, $p = ns$). Comme la taille de la fenêtre temporelle ne semble pas avoir d'importance, nous poursuivons nos analyses uniquement sur une fenêtre de 10 frames. Pour tester l'influence de la proximité des pics de saillance DESA sur la dispersion, une autre ANOVA à deux facteurs intra (la condition expérimentale V ou AV ; la dispersion moyennée sur toutes les frames de chaque plan ou sur les 10 frames suivant les pics de saillance de chaque plan) a été menée. Elle révèle un effet principal de la condition expérimentale ($F(1,126)=24.8$, $p < .001$), mais aucun effet des pics de saillance sonore ($F(1,126)=1.52$, $p = ns$).

Après les pics de saillance du modèle Energie Pour tester l'influence de la taille de la fenêtre d'analyse de la dispersion après les pics de saillance du modèle Energie, une ANOVA à deux facteurs intra (la condition expérimentale V ou AV ; la taille de la fenêtre temporelle 5, 10 ou 25 frames) a été menée. Elle révèle un effet principal de la condition expérimentale ($F(1,126)=39.6$, $p < .001$), conformément aux résultats de la section 2.1.4.1. Par contre, elle ne met en évidence aucun effet de la taille de la fenêtre temporelle ($F(2,124)=0.2$, $p = ns$). Comme pour les pics DESA présentés ci-dessus, nous poursuivons nos analyses uniquement sur une fenêtre de 10 frames. Pour tester l'influence de la proximité des pics de saillance du modèle Energie sur la dispersion, une autre ANOVA à deux facteurs intra (la condition expérimentale V ou AV ; la dispersion moyennée sur toutes les frames de

chaque plan ou sur les 10 frames suivant les pics de saillance de chaque plan) a été menée. Elle révèle un effet principal de la condition expérimentale ($F(1,126)=28.2$, $p < .001$), mais aucun effet des pics de saillance sonore ($F(1,126)=0.82$, $p = ns$).

Par souci de concision, nous ne présentons pas ici les résultats pour les autres métriques (durées de fixation, amplitudes de saccade, distance au centre, DKL). Elles suivent exactement le même modèle que la dispersion, à savoir des résultats équivalents à ceux de la section 2.1.4.1 pour l'effet de la condition expérimentale, et une absence d'effet des pics de saillance sonore.

2.2.4 Discussion

Dans cette section, nous avons posé la question de la nature de l'effet de l'information sonore sur l'exploration visuelle. Une importante modification sonore (un événement saillant) induit-elle forcément une importante modification de l'exploration visuelle ? Ou cet effet est-il de nature plus complexe ? Nous avons présenté deux modèles de saillance sonore, dans le but de comparer les résultats globaux obtenus lors de l'expérience 1, avec ceux obtenus sur les quelques frames suivant les événements sonores les plus saillants.

Pertinence des modèles de saillance sonore utilisés

Afin d'être sûr que les événements repérés par les modèles utilisés correspondent bien à une réalité perceptive, nous avons évalué le DESA et le modèle Energie. L'évaluation des modèles de saillance sonore demeure une question ouverte. Cette dernière est nettement plus problématique que pour les modèles de saillance visuelle, puisque nous ne disposons d'aucune mesure physique objective pour évaluer l'attention auditive. En effet, nous n'orientons que rarement nos oreilles vers les sources sonores d'intérêt ! Cependant, quelques rares méthodes ont été proposées (voir [Duangudom Delmotte 2012] et section 1.3.4). Ici, l'expérience que nous avons menée nous a permis de montrer que les deux modèles de saillance sonore que nous utilisons donnent des résultats davantage proches de la perception humaine qu'une génération aléatoire d'"événements sonores". De plus, nous avons mesuré une efficacité supérieure pour le modèle Energie que pour le modèle DESA, ce qui pourrait sembler surprenant au vu de la plus grande complexité de ce dernier. Cependant, il ne faut pas en tirer de conclusions hâtives. Le modèle Energie met davantage en avant les changements les plus évidents du signal auditif, lesquels ne sont pas nécessairement ceux qui attirent le plus l'attention. Par exemple, une petite voix dans un environnement bruité ne sera que difficilement repérée par le modèle Energie, alors qu'elle attirera grandement l'attention. A l'inverse, le DESA sera plus à même de repérer cette voix, grâce notamment à l'élimination du bruit environnant

par le banc de filtre de Gabor, et à l'attribut fréquentiel FIM [Evangelopoulos & Maragos 2006]. Un défaut majeur de l'évaluation que nous avons mise en place est que pour l'effectuer, les participants ont besoin d'écouter chaque bande-son plusieurs fois (au moins une fois par jeu de pics). Or, la répétition des écoutes peut biaiser la perception de la saillance par les auditeurs : seuls les changements les plus évidents demeurent saillants, favorisant donc les pics du modèle Energie. Toutefois, étant donné que les pics du DESA et du modèle Energie sont jugés nettement plus saillants que les pics aléatoires, les analyses subséquentes sont légitimes.

Evénements sonores et exploration visuelle

Une loi fondamentale de l'intégration multimodale stipule que l'intégration est plus probable et plus forte lorsque les stimuli issus des différentes modalités sont simultanés [Meredith & Stein 1986, Meredith *et al.* 1987, Stein & Meredith 1993]. Nous avons fait l'hypothèse qu'à proximité des événements sonores les plus saillants, l'interaction entre ces derniers et l'information visuelle correspondante sera plus forte, et donc que l'exploration visuelle sera davantage modifiée entre les conditions Visuelle et AudioVisuelle. De plus, il a été montré que la largeur de la fenêtre temporelle d'intégration varie considérablement selon la nature des stimuli en présence (entre 20 et 400 ms) [Vatakis & Spence 2006, Recanzone 2009]. Pour tester cette hypothèse, nous avons calculé la valeur moyenne des caractéristiques des mouvements oculaires à l'intérieur de fenêtres temporelles de différentes tailles, suivant chaque pic de saillance repéré par les modèles.

Nous n'avons trouvé aucune différence entre ces valeurs et celles calculées de manière globale sur l'ensemble des frames de chaque plan, quelle que soit la taille de la fenêtre d'analyse (20 ms, 400 ms ou 1 s). Cette absence de différence indique que la nature de l'interaction entre la saillance sonore et la saillance visuelle n'est pas linéaire, et que d'autres paramètres doivent être pris en considération. Il est aussi possible que l'interaction audio-

Psychocinématique 1

"Au début, vous avez un élément, c'est l'image. L'image, on peut la comparer à un diamant. Le diamant brut, c'est une pierre d'une beauté extraordinaire, c'est la lumière. Puis les gens apprennent à polir le diamant [...] Mais il y a une unité totale dans ce procédé, parce que c'est un *élément. Le parlant, c'est* deux *éléments : le son, et l'image. Sur le plan de la matière simple, facile à comprendre, c'est de la céramique : vous avez la terre glaise, vous avez la couleur, vous mettez le tout dans un four, et ça devient une céramique, ça devient un autre élément unitaire. Le drame du cinema, c'est que très peu de gens arrivent à faire cette unité."*

Henri Langlois,
fondateur de la cinémathèque française

visuelle s'exprime sur de plus grandes périodes de temps, par exemple à l'échelle du plan entier. En effet, pour apprécier une pièce musicale, il est nécessaire d'en percevoir sa structure mélodique globale, même si les auditeurs n'en sont que rarement directement conscients [Dowling *et al.* 1995]. De la même manière, pour comprendre la façon dont l'interaction audiovisuelle agit lors de l'exploration d'une scène naturelle, les relations complexes entre les objets sonores et visuels, la "structure" globale (sur le long ou le moyen terme) de la scène audiovisuelle pourraient jouer un rôle important (voir l'encart Psychocinématique 1).

Dans la discussion de la section précédente, nous avons rappelé que la plupart des études portant sur l'intégration audiovisuelle utilisent des stimuli simples et artificiels. Ce faisant, les auteurs se concentrent sur les interactions bas niveau entre les deux modalités, espérant limiter les effets de plus haut niveau. Dans les scènes audiovisuelles plus complexes telles que celles qui nous intéressent, les effets de haut niveau prennent de l'importance, au point de jouer un rôle prépondérant dans l'intégration audiovisuelle. Un bon exemple de l'expression de ces effets de haut niveau est offert par la psychomusicologie, et son approche cognitive du rôle de la musique dans la perception, l'appréciation, ou la mémorisation des images [Bolivar *et al.* 1994, Boltz 2004, Boltz *et al.* 2009, Cohen 2005, Cohen 2014]. Certaines études issues de cette discipline suggèrent que la perception de la musique est modulée par de nombreuses variables, et qu'un même morceau pourra évoquer de multiples réactions d'un auditeur à l'autre selon l'heure, le contexte social ou encore le lieu dans lequel il est écouté [Sloboda & O'Neill 2001]. De plus, le vécu de chacun joue beaucoup sur les émotions ou les associations intermodales suscitées par une scène audiovisuelle [Juslin & Laukka 2004]. Dans [Gabrielsson 2001], l'auteur introduit le concept de *Strong Experience in Music* (SEM) et analyse les témoignages de près de 900 personnes décrivant un épisode musical de leur vie les ayant profondément bouleversés. Gabrielsson propose une analyse en composantes principales des facteurs expliquant ce que l'instant musical évoqué avait de si spécial. Des facteurs aussi variés que la beauté transcendante de la musique, des réactions physiologiques (chair de poule, respiration accélérée), une synergie cathartique avec les musiciens (ils ont compris ce que je ressens et arrivent à l'exprimer), une altération de l'espace et du temps, une atmosphère spéciale ressentie à l'unisson par le public... ont été évoqués par les participants, et illustrent l'extrême complexité de la perception musicale. Cette grande diversité dans la perception musicale se transcrit directement dans la perception audiovisuelle : de la même manière qu'un ensemble d'objets musicaux (notes, accords, mélodies) n'est pas perçu à l'unisson par chacun, un ensemble d'objets multimodaux (sons et images) ne sera pas non plus interprété de la même façon, et n'induira pas forcément les réactions homogènes que nous cherchions à mesurer.

Il va de soi que ces facteurs de haut niveau, dépendant du vécu, de la mémoire à long terme de chacun ne sont pas aisés à modéliser. Cependant depuis quelques années se développent des modèles d'attention visuelle prenant en compte

certains processus descendants tels que l'anticipation, la récompense, la connaissance préalable du contexte, la tâche demandée... [Torralba *et al.* 2006, Zhang *et al.* 2008, Li *et al.* 2010]. La plupart de ces modèles s'appuient sur le formalisme bayésien, qui propose de modéliser la probabilité d'occurrence d'un événement dans un contexte donné à la fois en fonction de la situation actuelle, et des connaissances préalables que l'on a du contexte. Cette approche a également été utilisée pour modéliser l'intégration multimodale [Ernst & Banks 2002, Ernst & Bülthoff 2004, Burr & Alais 2006]. Ici, l'hypothèse est que l'humain combine les informations multimodales issues de son environnement, en associant percepts sensoriels et connaissances préalables (*prior knowledge*), et en les pondérant en fonction de leur fiabilité (*i.e.* de l'inverse de leur variance). Ainsi, un axe de recherche prometteur consisterait à coupler ces applications attentionnelles et multisensorielles du formalisme bayésien, en proposant un modèle de saillance audiovisuelle prenant en compte certaines relations de haut niveau entre objets visuels et objets auditifs.

2.3 Conclusion

Au cours de ce chapitre, nous avons évalué l'influence de l'information sonore sur l'exploration libre de scènes dynamiques au contenu audiovisuel varié. Nous nous sommes basés sur le plan comme unité temporelle de base, et avons décomposé son exploration en quatre phases. Nous avons montré que la suppression de la bande-son entraîne une augmentation de variabilité entre les positions oculaires de différents observateurs ainsi qu'une diminution de l'amplitude de leurs saccades. De plus, les conditions expérimentales affectent les cartes de densité de position oculaire : deux personnes dans la même condition regardent davantage les mêmes régions que deux personnes dans des conditions différentes. Pour sonder la nature des interactions entre les informations sonore et visuelle, nous avons ensuite comparé les mouvements oculaires moyens avec ceux effectués juste après des événements sonores particulièrement saillants. Afin de repérer de tels événements dans les bandes-son de nos vidéos, nous avons décrit et utilisé deux modèles de saillance sonore. Nous n'avons trouvé aucune différence consécutive à la proximité de tels événements. Ceci met en défaut notre hypothèse, et suggère que les relations entre saillance sonore et saillance visuelle sont de nature plus complexes.

Toutefois, il est important de considérer que les résultats de cette expérience ont été mesurés à partir de vidéos au contenu audiovisuel très varié. Les interprétations énoncées dans les discussions de ce chapitre sont donc d'ordre général et pourraient se révéler fausses si l'on restreignait l'analyse à un type d'information particulier. Par exemple, de nombreuses études ont montré que la perception des visages suit des processus tout à fait particuliers, et est bien différente de la perception d'objets moins sociaux (voir chapitre 4).

De même, le concept de "saillance sonore" n'est pas forcément pertinent pour tous les types de scènes auditives. Ainsi, s'il est facile d'identifier les pics de saillance sonore lors d'un échange de coups de feu, cet exercice devient nettement plus subtil pour une scène de paysage, ou seul un vent continu se fait entendre. Il serait donc logique de différencier les relations multimodales en fonction du type de scènes audiovisuelles, et de comparer les interactions entre différents profils visuels et différents profils sonores.

Chapitre 3

Effets de différents contenus sonores sur différentes catégories visuelles

Sommaire

Dans ce troisième chapitre, nous souhaitons approfondir les résultats obtenus au chapitre précédant en contrôlant le contenu audiovisuel de nos stimuli. Nous présentons l'expérience 2, au cours de laquelle les mouvements oculaires de 72 nouveaux participants ont été enregistrés alors qu'ils regardaient des vidéos classées selon quatre catégories visuelles et quatre conditions expérimentales manipulant le contenu sonore. Dans un premier temps, nous présentons notre protocole et décrivons les différences observées sur les mouvements oculaires, tant entre les catégories visuelles qu'entre les conditions expérimentales. Les modèles de saillance sonore présentés au chapitre précédent sont également utilisés afin de quantifier l'effet des événements sonores sur les mouvements oculaires. Dans un second temps, nous présentons deux méthodes de modélisation statistique, que nous utilisons pour comprendre quels attributs visuels (saillance statique, saillance dynamique, biais de centralité...) expliquent le mieux les mouvements oculaires dans les différentes catégories visuelles et conditions expérimentales.

3.1 Introduction

Nous nous intéressons à la manière dont les informations visuelle et sonore d'une scène dynamique interagissent pour guider notre attention. De précédentes études, basées sur des scènes naturelles statiques présentées avec des sons localisés suggèrent que les deux informations unimodales sont intégrées linéairement [Onat *et al.* 2007, Quigley *et al.* 2008]. Cependant, d'autres travaux indiquent que la nature de l'intégration n'est pas universelle, mais dépend très fortement du contenu des stimuli sonores et visuels. Une équipe d'architectes paysagistes a demandé à 75 participants de juger le plaisir qu'ils avaient à percevoir 36 combinaisons son / images "naturelles" [Carles *et al.* 1999]. Ils ont constaté que les notes attribuées à une image donnée variaient systématiquement avec les sons présentés simultanément, et que les stimuli jugés comme étant les plus agréables étaient ceux dont les informations des deux modalités étaient congruentes (par exemple un ruisseau avec le bruit de l'eau, ou un village avec son ambiance sonore). Une autre équipe de recherche a tenté de quantifier ces résultats qualitatifs en enregistrant les mouvements oculaires de huit participants visionnant un même extrait vidéo dans quatre conditions expérimentales : avec la bande-son originale, muet, avec des sons d'armes à feu ou avec des pleurs de bébé [Vilaró *et al.* 2012]. Si les cartes de densité des positions oculaires des différents sujets se ressemblent grandement d'une condition expérimentale à l'autre, certaines différences ont été constatées. Toutefois, le faible nombre de sujets et l'unique stimulus visuel utilisé empêchent les auteurs d'interpréter ou de généraliser les différences observées.
Afin de comprendre plus précisément comment les caractéristiques physiques des signaux sonores et visuels interagissent lors de la perception audiovisuelle, certains auteurs ont utilisé un paradigme de jugement temporel de la synchronie entre le son et l'image de scènes plus ou moins complexes [Vatakis & Spence 2006]. Des vidéos présentant par exemple des visages en train de parler, ou des objets en mouvement (comme un marteau frappant une table) ont été présentées avec différents décalages temporels entre les deux modalités. Les participants ont jugé plus précisément l'ordre relatif d'apparition des signaux auditif et visuel pour les objets en mouvement que pour les visages parlants. Ceci peut signifier qu'il est plus facile de discriminer temporellement son et image pour des stimuli audiovisuels de faible complexité (son bref et geste unique) que pour des stimuli dont les propriétés varient de manière plus continue (sons de parole riches et gestes articulatoires complexes). D'autres auteurs ont émis l'idée que puisqu'un son de parole présente une corrélation temporelle très fine avec le visage du locuteur, juger l'ordre temporel d'apparition entre son et image peut être plus difficile que pour des scènes audiovisuelles plus abruptes, et donc moins temporellement corrélées : c'est le phénomène de ventriloquie temporelle [Vroomen & Stekelenburg 2011]. Ainsi, la nature de l'intégration audiovisuelle semble être liée, au moins en partie, aux propriétés physiques du contenu audiovisuel. Une récente étude s'intéressant à la catégorisation de scènes audiovisuelles naturelles (intérieure, extérieure, calme, vivante...) va dans le sens de cette

idée [Rummukainen *et al.* 2014]. Les auteurs ont présenté 19 vidéos représentant différentes scènes naturelles à des participants au moyen d'un système de projection particulièrement immersif (trois projecteurs produisant une image de 226°de champ horizontal, résolution de 4320 × 1080 pixels, 29 hauts-parleurs placés de manière à assurer une spatialisation sonore haute-fidélité). Une analyse statistique des propriétés des stimuli couplée aux retours subjectifs des participants a permis de déterminer les attributs audiovisuels les plus utilisés pour catégoriser une scène dans un groupe plutôt que dans un autre. Il s'agit du mouvement, de la bruyance et l'animation de la scène. Cependant, les mouvements oculaires n'ont pas été enregistrés (les auteurs travaillent actuellement sur l'inclusion d'un oculomètre dans leur dispositif), et les contributions individuelles des différents attributs dans la saillance audiovisuelle n'ont donc pu être précisément déterminées.

Hypothèses

Ici, nous formulons l'hypothèse que l'effet du son sur l'exploration visuelle ne s'exprime pas de la même façon selon le type d'association audiovisuelle. Au sein d'une catégorie visuelle donnée, nous regardons si différentes conditions expérimentales modulent l'exploration visuelle, et si cette modulation est accentuée après un événement sonore saillant. Nous définissons quatre catégories au contenu visuel différent, et les présentons avec leur bande-son originale, avec la bande-son d'une autre vidéo appartenant à la même catégorie visuelle, ou avec la bande-son d'une vidéo appartenant à une catégorie visuelle différente. Nous faisons l'hypothèse que plus la modalité sonore contient une information riche, plus elle aura une influence sur l'exploration de cette dernière. A l'inverse, un contenu sonore sans événement particulièrement saillant pourrait n'avoir aucune influence sur l'exploration d'une scène visuelle complexe. De plus, l'exploration pourrait être différemment affectée selon le degré d'incongruence de la bande-son avec l'information visuelle : si les événements sonores sont en rapport direct avec les événements visuels, ils pourraient renforcer la saillance de ces derniers.

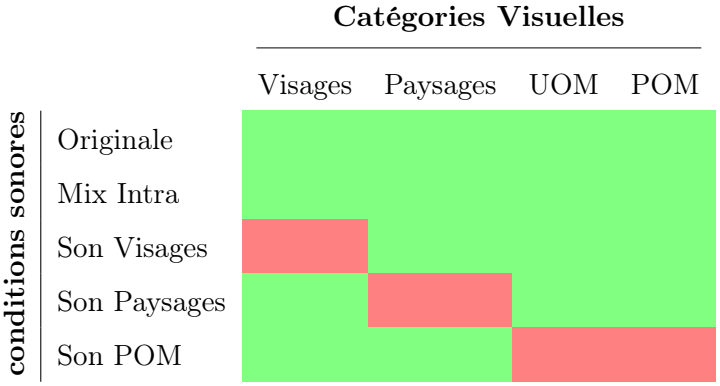
3.2 Expérience 2

3.2.1 Design expérimental

3.2.1.1 Participants

72 personnes ont participé à l'expérience : 41 hommes et 31 femmes, âgés entre 20 et 35 ans ($M = 24.3\,; SD = 4.6$). Les participants étaient naïfs quant au but de l'expérience, et avaient pour consigne de regarder librement et attentivement les vidéos présentées. Tous les participants étaient de langue maternelle française et avaient une vue normale ou corrigée à la normale. Aucun n'a reporté de trouble

Table 3.1 – Association des catégories visuelles et des conditions expérimentales. Quatre conditions expérimentales sont définies pour chaque catégorie visuelle (cases vertes). Comme le contenu des bandes-son des catégories POM et UOM sont assez proches, nous n'avons pas créé de condition Son UOM.



auditif. Chacun a donné son consentement éclairé à prendre part à l'expérience.
Le dispositif et l'organisation des données sont rigoureusement les mêmes que ceux utilisés lors de l'expérience 1 et décrits section 2.1.2.2.

3.2.1.2 Stimuli

Nous avons utilisé 60 vidéos, classées *a priori* selon quatre catégories visuelles :
15 vidéos **Visages** (de 2 à 4 personnes ayant une conversation),
15 vidéos **Un Objet en Mouvement** (voiture de police, maquette d'avion...),
15 vidéos **Plusieurs Objets en Mouvement** (réaction en chaîne, flipper...) et
15 vidéos **Paysages** (bords de mer, prairies...).
Chaque vidéo ne contient qu'un seul plan.
Une description accompagnée d'une frame illustrative de chacune des scènes utilisées est disponible en Annexe B. Notre choix s'est porté sur ces catégories car elles sont représentatives de la plupart des scènes dont nous sommes témoins dans la vie de tous les jours, et car chacune possède un profil de saillance visuelle bien différent. En effet, les visages et le mouvement étant connus comme les principaux attributs visuels attirant l'attention [Yarbus 1967, Hershler & Hochstein 2005, Mital *et al.* 2010], les trois premières catégories présentent des distributions de saillance parcimonieuses, avec un objet saillant "principal" caractérisé par l'unique objet en mouvement pour la catégorie Un Objet en Mouvement (UOM), ou autant d'objets saillants "principaux" qu'il y a d'objets en mouvement ou de visages pour les catégories Plusieurs Objets en Mouvement (POM) ou Visages. A l'inverse, les vidéos de la catégorie Paysages ont un profil de saillance plus uniforme, aucune région n'attirant particulièrement le regard.
Le profil de saillance sonore des bandes-son associées aux vidéos change aussi beau-

coup d'une catégorie à l'autre. Dans les catégories UOM et POM, l'information sonore est principalement constituée des sons ponctuels des objets en mouvement (par exemple le son d'un marteau sur une pierre). Dans la catégorie Visages, il s'agit du son de parole des différents locuteurs en présence, les dialogues étant tous en français. Enfin, dans la catégorie Paysages, il s'agit de sons plus continus et répétitifs (bruit du vent, de la pluie, de la mer...). Dans chaque catégorie, toutes les sources sonores sont présentes à l'image.

Les vidéos ont une résolution de 720×576 pixels (30×24 degrés d'angle visuel), une fréquence de 25 images par seconde, et durent entre 10 s et 24.8 s ($M = 16.9\,; SD = 4.8$). Les bandes-son sont monophoniques et échantillonnées à 48000 Hz. Si à l'origine, le son était stéréophonique, nous avons ajouté les deux canaux et envoyé la somme dans chaque écouteur.

Pour étudier l'influence du son sur l'exploration visuelle en fonction du type d'association son/image, nous avons créé quatre versions de chaque vidéo (correspondant à quatre conditions expérimentales), chacune avec une bande-son différente (Table 3.1). Pour les vidéos de la catégorie

- Visages : chaque vidéo possédait une version originale (condition **Originale**), une version avec la bande-son d'une autre vidéo de la catégorie Visages (condition **Mix Intra**), une version avec la bande-son d'une vidéo de la catégorie POM (condition **Son POM**) et une version avec la bande-son d'une vidéo de la catégorie Paysages (condition **Son Paysages**).
- Paysages : chaque vidéo possédait une version originale (condition **Originale**), une version avec la bande-son d'une autre vidéo de la catégorie Paysages (condition **Mix Intra**), une version avec la bande-son d'une vidéo de la catégorie Visages (condition **Son Visages**) et une version avec la bande-son d'une vidéo de la catégorie POM (condition **Son POM**).
- UOM : chaque vidéo possédait une version originale (condition **Originale**), une version avec la bande-son d'une autre vidéo de la catégorie UOM (condition **Mix Intra**), une version avec la bande-son d'une vidéo de la catégorie Visages (condition **Son Visages**) et une version avec la bande-son d'une vidéo de la catégorie Paysages (condition **Son Paysages**).
- POM : chaque vidéo possédait une version originale (condition **Originale**), une version avec la bande-son d'une autre vidéo de la catégorie POM (condition **Mix Intra**), une version avec la bande-son d'une vidéo de la catégorie Visages (condition **Son Visages**) et une version avec la bande-son d'une vidéo de la catégorie Paysages (condition **Son Paysages**).

Les affectations des différentes bandes-son aux différentes vidéos ont été faites de manière aléatoire. Lorsque bande-son et vidéo n'avaient pas la même longueur, nous avons tronqué le plus grand des deux. Au total, nous sommes donc en présence de 4 catégories visuelles × 15 vidéos × 4 bandes-son = 240 stimuli différents. Nous les avons pseudo-aléatoirement répartis en 4 jeux de 60 stimuli comportant 15 vidéos de chaque catégorie visuelle, équitablement réparties entre les conditions expérimentales. Par exemple, dans un de ces 4 jeux, dans la catégorie Visages, il y a 4 vidéos dans la condition Originale, 4 vidéos dans la condition Mix Intra, 4 vidéos dans la

condition Son POM et 3 vidéos dans la catégorie Son Paysages.

3.2.1.3 Protocole

Chaque participant a visionné un des 4 jeux de 60 stimuli décrits ci-dessus. Le protocole était identique à celui de l'expérience 1, illustré Figure 2.1b. Chaque expérience était précédée par une procédure de calibration, durant laquelle les participants devaient concentrer leur regard sur 9 cibles réparties sur une grille 3×3 occupant tout l'écran. Une correction de la dérive du regard était effectuée entre chaque vidéo et une nouvelle calibration était effectuée toutes les 15 vidéos, ou si la dérive excédait 0.5 degré. Afin d'éviter tout effet d'ordre ou de fatigue, les vidéos étaient présentées dans un ordre aléatoire. Une pause était systématiquement proposée toutes les 15 vidéos, et les participants étaient informés qu'ils pouvaient se reposer à n'importe quel moment entre deux vidéos. Le cas échéant, une calibration était à nouveau effectuée à la reprise de l'expérience. Une expérience durait environ 25 minutes. Au final, chaque vidéo a été vue par $72/4 = 18$ participants différents dans chaque condition expérimentale.

3.2.2 Résultats

Pour comparer les différents types d'explorations enregistrés, nous utilisons les métriques introduites au chapitre précédent (section 2.1.3). Dans un premier temps, nous présentons les différences induites par le contenu visuel, toutes conditions expérimentales confondues. Dans un second temps, nous nous intéressons aux interactions entre contenus visuels et sonores. Des corrections de Bonferroni ont été appliquées afin d'ajuster les niveaux de significativité lors de comparaisons multiples (les différences ont été considérées significatives pour $p < \frac{.05}{n}$ avec n le nombre de comparaisons effectuées).

3.2.2.1 Selon la catégorie visuelle

Nous présentons les différences induites par le contenu visuel, toutes conditions expérimentales confondues. Pour chaque métrique, nous donnons les résultats obtenus en moyenne sur l'ensemble des vidéos (analyse globale) ainsi qu'en fonction du temps (analyse temporelle).

Analyse globale...

Comme l'indiquent les Figures 3.1 et 3.2, nous observons une forte variation des différentes métriques selon la catégorie visuelle. Nous avons mené une ANOVA à un facteur inter (les catégories visuelles) sur les valeurs de chaque métrique. Afin de ne comparer que les différences dues au contenu visuel, nous prenons en compte

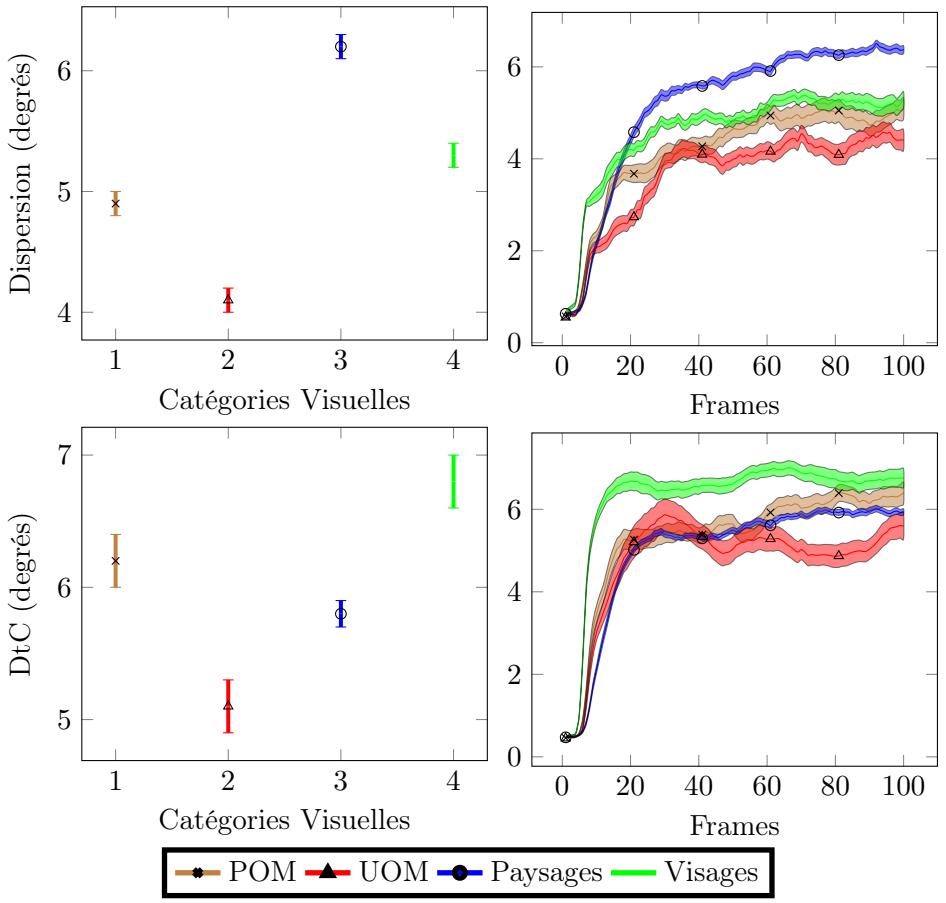


Figure 3.1 – **Gauche :** dispersion et distance au centre (DtC) moyennées sur l'ensemble des frames de chaque stimulus, toutes conditions expérimentales confondues : 4 × 15 items par catégorie visuelle. **Droite :** évolutions temporelles de la dispersion et de la distance au centre moyennées sur chaque stimulus, toutes conditions expérimentales confondues. Les valeurs sont données en degrés angulaires, les barres d'erreur correspondent aux erreurs standards.

dans chaque catégorie les résultats de toutes les conditions expérimentales. Ainsi, pour la dispersion et la distance au centre, chaque niveau a 60 items (15 stimuli × 4 bandes-son). Pour les amplitudes de saccade et les durées de fixation, chaque niveau a 288 items (72 sujets × 4 bandes-son).

... et analyse temporelle

Il existe également une certaine évolution des différentes métriques en fonction du temps. Dans chaque catégorie, l'évolution temporelle a une allure semblable à celle décrite au chapitre précédent, section 2.1.4.2 : une latence durant les 5 frames suivant l'apparition du stimulus, suivie d'une augmentation rapide, puis d'une stabilisation autour d'une valeur moyenne, plus ou moins élevée selon les catégories visuelles. Pour les amplitudes de saccade et les durées de fixation, cette stabilisation n'est pas évidente : on observe plutôt une légère décroissance, du moins dans

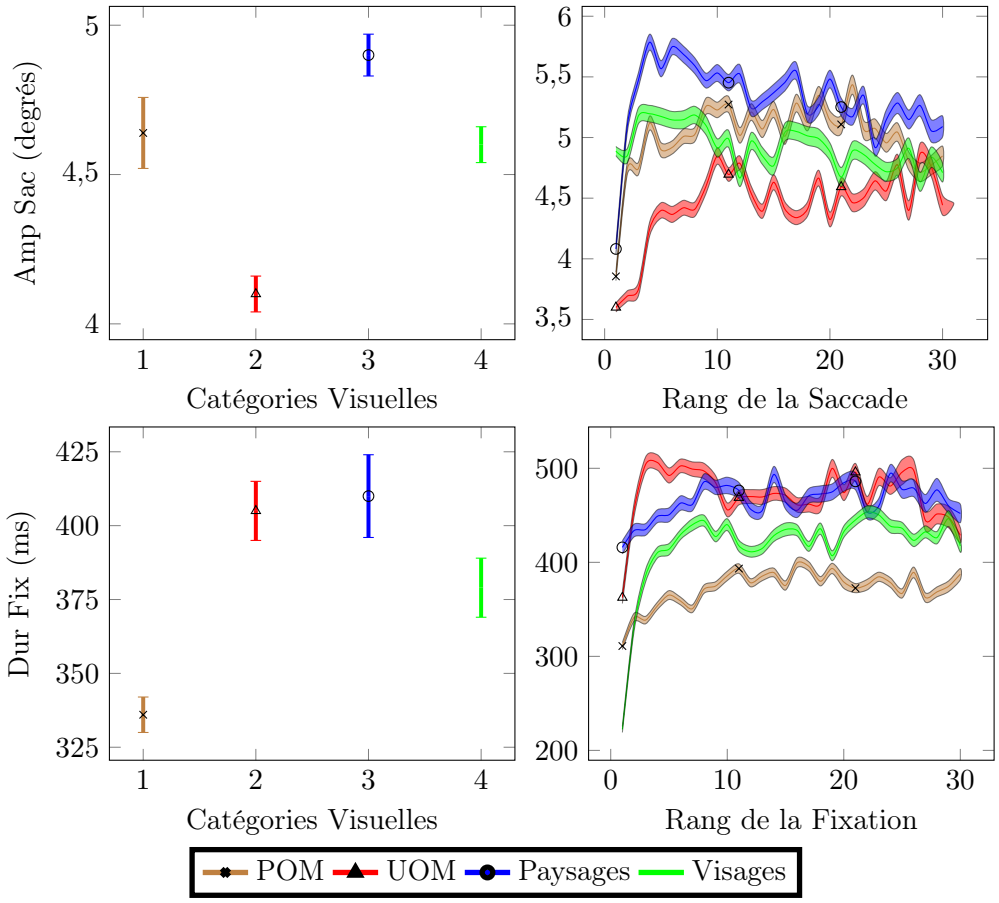


Figure 3.2 – Gauche : amplitudes de saccade et durées de fixation médianes de l'ensemble des saccades et fixations de chaque sujet, moyennées sur toutes les conditions expérimentales : 4 x 72 items par catégorie visuelle. **Droite :** amplitudes de saccade et durées de fixation en fonction de leur rang, moyennées sur chaque sujet, toutes conditions expérimentales confondues. Les barres d'erreur correspondent aux erreurs standards.

certaines catégories visuelles. Pour quantifier et comparer cette évolution selon les catégories visuelles, nous effectuons des ANOVA à un facteur inter : la catégorie visuelle, et un facteur intra : le temps. Pour la dispersion et la distance au centre, le facteur "temps" possède 10 niveaux correspondant aux valeurs moyennes des métriques entre les frames 1 et 10, 10 et 20, ..., 90 et 100, ce qui nous permet une analyse fine de leur évolution temporelle. Chaque niveau possède 60 items (15 stimuli $\times$ 4 bandes-son). Pour les amplitudes de saccade et les durées de fixation, le facteur "temps" possède 5 niveaux correspondant aux valeurs moyennes des métriques entre les rangs de saccade (ou de fixation) 1 et 5, 5 et 10, ..., 20 et 25. Chaque niveau possède 288 items (72 sujets x 4 bandes-son).

Dispersion

Il existe un effet principal de la catégorie visuelle ($F(3,177) = 41.5$, $p < .001$).

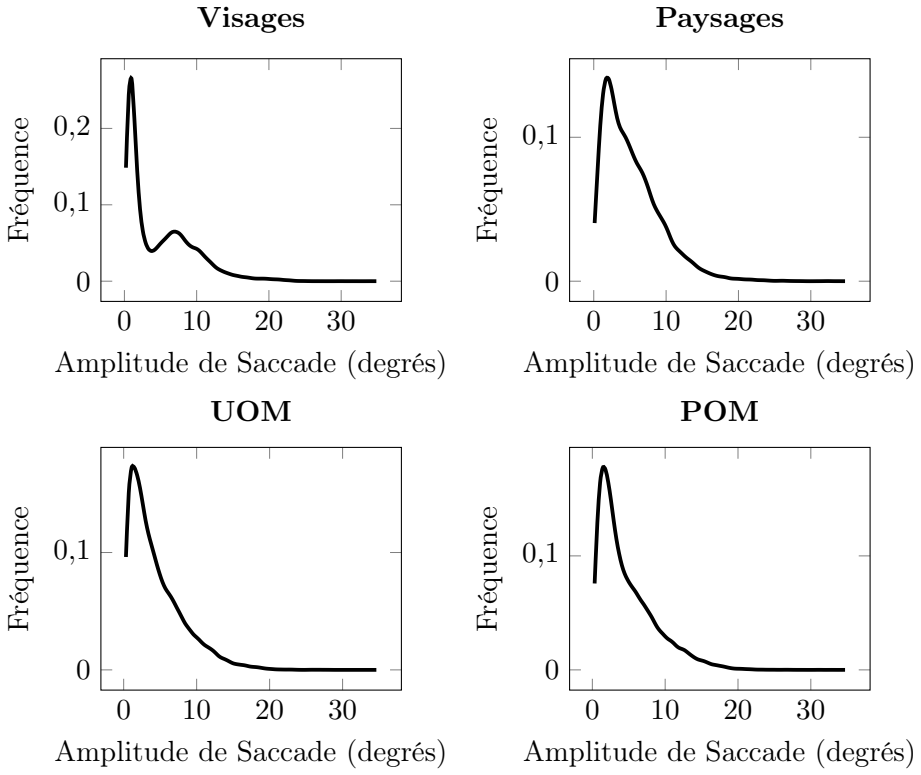


Figure 3.3 – Distributions des amplitudes de saccade dans les quatre catégories visuelles, toutes conditions expérimentales confondues. Les densités de probabilité ont été estimées sur 100 points régulièrement espacés couvrant l'ensemble des données (fonction Matlab ksdensity).

La dispersion de la catégorie Paysages est supérieure à toutes les autres (tous les $p < .001$), et celle de la catégorie UOM est inférieure à toutes les autres (tous les $p < .001$). La dispersion de la catégorie POM n'est pas significativement différente de celle de la catégorie Visages ($p = .19$).
Il existe également un effet du temps ($F(9,531) = 393.7$, $p < .001$). La dispersion augmente rapidement entre les trois premiers niveaux (frames 1à 10, 10 à 20 et 20 à 30, tous les $p < .001$), puis se stabilise doucement. Il n'y a plus de différence significative entre les 4 derniers niveaux (tous les $p = 1$).
L'interaction est également significative ($F(27,1593) = 7.4$, $p < .001$), cette interaction est notamment due à la lente stabilisation de la catégorie Paysages par rapport aux autres catégories visuelles. En effet, alors que pour les autres catégories, on ne constate plus aucune différence à partir du 3ème niveau temporel, il faut attendre le 5ème niveau pour les Paysages.

Distance au centre

Il existe un effet principal de la catégorie visuelle ($F(3,177) = 17.3$, $p < .001$). La distance au centre de la catégorie UOM est inférieure à toutes les autres (UOM vs.

Paysages $p = .03$, UOM vs. Visages et POM $p < .001$). La distance au centre de la catégorie POM n'est pas significativement différente ni de celle dans la catégorie Visages ($p = .11$), ni de celle dans la catégorie Paysages ($p = .47$). Enfin, la distance au centre dans la catégorie Visages est supérieure à celle dans la catégorie Paysages ($p < .001$).
Il existe également un effet du temps ($F(9,531) = 403.3$, $p < .001$). La distance au centre augmente rapidement entre les deux premiers niveaux (tous les $p < .001$), puis se stabilise. Il n'y a plus de différence significative entre les 5 derniers niveaux (tous les $p = 1$).
L'interaction est également significative ($F(27,1593) = 6.3$, $p = .001$), également due aux vitesses de stabilisation différentes selon les catégories : la catégorie Visages se stabilise la première, dès le $2^{\text{ème}}$ niveau, alors qu'il faut attendre le $3^{\text{ème}}$ pour UOM et Paysages, et le $6^{\text{ème}}$ pour POM.

Amplitudes de saccade

Il existe un effet principal de la catégorie visuelle ($F(3,861) = 53.2$, $p < .001$). Les amplitudes de saccade médianes des quatre catégories visuelles sont toutes différentes les unes des autres (tous les $p < .001$). Les plus grandes sont celles de la catégorie Paysages, suivies de celles des catégories Visages, POM, et UOM. La Figure 3.3 montre que pour chaque catégorie visuelle, les amplitudes de saccade suivent une distribution asymétrique positive avec un mode principal autour de 1.5°, comme c'était le cas lors de l'expérience 1 (Figure 2.3). Cependant certaines différences apparaissent, reflétant les différences constatées entre les moyennes. Il y a davantage de saccades de moyennes amplitudes (entre 3°et 10°) dans la catégorie Paysages que dans les autres. A l'inverse, le premier mode de la catégorie Visages est plus prononcé, signe qu'elle contient davantage de saccades de petites amplitudes. De plus, dans cette catégorie uniquement, un second mode apparait autour de 7°.
Il existe également un effet du temps ($F(4,1148) = 23.9$, $p < .001$). Les amplitudes de saccade du premier niveau (rangs 1 à 5) sont inférieures à toutes les autres (tous les $p < .001$). On observe une légère diminution à partir du $2^{\text{ème}}$ niveau, mais seul le $5^{\text{ème}}$ niveau est significativement inférieur à ce dernier ($p = .001$).
L'interaction est également significative ($F(12,3444) = 11.0$, $p < .001$). En effet, alors que les catégories UOM et POM se stabilisent dès le $2^{\text{ème}}$ niveau (tous les $p = 1$), les amplitudes de saccade des catégories Visages et Paysages décroissent légèrement : le $5^{\text{ème}}$ niveau est inférieur au $2^{\text{ème}}$ ($p < .005$).

Durées de fixation

Il existe un effet principal de la catégorie visuelle ($F(3,861) = 50.5$, $p < .001$). Les durées de fixation médianes les plus longues sont celles des catégories UOM et Paysages (pas de différence entre elles deux, $p = 1$). Viennent ensuite celles de la catégorie Visages, puis POM (tous les autres $p < .001$). Nous n'avons pas représenté les distributions, car leur forme ne change guère d'une catégorie à l'autre, et sont

semblables à celles représentées lors de l'expérience 1, Figure 2.3.
Il existe également un effet du temps ($F(4,1148) = 41.9$, $p < .001$). Les durées de fixation du premier niveau sont inférieures à toutes les autres (tous les $p < .001$). Aucun des autres niveaux ne diffère (tous les $p = 1$).
L'interaction est également significative ($F(12,3444) = 5.5$, $p < .001$). Pour la catégorie POM, l'augmentation initiale est nettement moins rapide : les deux premiers niveaux ne sont pas différents l'un de l'autre ($p = .42$). A l'inverse, la catégorie UOM se stabilise si rapidement que même le premier niveau n'est pas différent des autres (entre niveau 1 et 2, $p = .70$, tous les autres $p = 1$).

Conformément aux hypothèses formulées, nous observons un effet principal de la catégorie visuelle pour toutes les métriques que nous utilisons pour caractériser l'exploration visuelle. Certaines des différences constatées entre les catégories visuelles s'interprètent assez facilement. Il est logique que la dispersion soit la plus grande dans la catégorie Paysages, dans la mesure où son contenu visuel ne présente aucune région attirant particulièrement le regard. N'ayant rien pour diriger leur attention, les observateurs explorent l'ensemble de la scène en ordre dispersé, par de grandes saccades et de longues fixations. A l'inverse, la dispersion est la plus faible dans la catégorie UOM car un seul objet d'intérêt attire tous les regards, provoquant de petites saccades et de longues fixations (sur l'objet en question). De plus, cet unique objet en mouvement est souvent au centre de l'écran, ce qui explique la faible distance au centre. Par contre, les catégories Visages et POM induisent des explorations visuelles moins évidentes à caractériser, sans doute à cause de la plus grande complexité qui les compose. La catégorie Visages est particulièrement différente des autres, avec une distribution d'amplitude de saccade bimodale. Nous l'étudierons en détail au Chapitre 4. L'évolution temporelle des métriques est semblable à celle constatée au chapitre précédent : un biais causé par la croix de fixation au début de l'exploration, une variation rapide suivie d'une stabilisation. Nous constatons également que cette dernière phase met plus ou moins de temps à s'établir selon les catégories : alors qu'elle intervient très rapidement pour la catégorie UOM, elle met plus de temps pour la catégorie Paysages. A présent, intéressons nous à l'effet de différents contenus sonores sur chacune de ces catégories.

3.2.2.2 Selon l'association audiovisuelle

Dans cette section, nous analysons pour chaque catégorie visuelle l'effet de la bande-son. Nous testons cet effet en comparant les mouvements oculaires enregistrés dans les différentes conditions expérimentales (Originale, Mix Intra, Son POM, Son Paysages, Son Visages), en moyenne sur toute la vidéo, puis localement, en moyennant sur les quelques frames suivant un événement sonore.

En moyenne

Pour chaque catégorie visuelle et chaque métrique, nous avons mené une ANOVA à un facteur intra (la condition expérimentale). Pour la dispersion et la distance au centre, chaque niveau a 15 items (un par stimulus). Pour les amplitudes de saccade et les durées de fixation, chaque niveau a 72 items (un par participant). Nous ne reportons ici que les effets significatifs. Pour une présentation détaillée de l'analyse, se référer à l'Annexe C.1.
La seule catégorie visuelle à être affectée par les conditions expérimentales est la catégorie **Visages**, dans laquelle la dispersion est plus faible et les saccades plus courtes dans la condition Originale que dans toute autre condition.

- **Dispersion :** Il existe un effet principal de la condition expérimentale ($F(3,42) = 17.97$, $p < .001$). La dispersion dans la condition Originale est inférieure à toutes les autres ($4.8° \pm 0.3$, tous les $p < .001$). Il n'existe pas de différence significative entre les autres conditions expérimentales (Son POM ($5.6° \pm 0.3$) vs. Mix Intra ($5.3° \pm 0.2$), $p = .07$; Son Paysages ($5.5° \pm 0.3$) vs. Son POM ($5.6° \pm 0.3$), $p = 1$; Son Paysages ($5.5° \pm 0.3$) vs. Mix Intra ($5.3° \pm 0.2$), $p = .79$).
- **Amplitude de saccade :** Il existe un effet principal de la condition expérimentale ($F(3,213) = 6.2$, $p < .001$). Les moyennes des amplitudes de saccade médianes dans la condition Originale sont inférieures à toutes les autres ($4.3° \pm 0.1$, tous les $p < .01$). Il n'existe pas de différence significative entre les autres conditions expérimentales (Son Paysages : $4.6° \pm 0.1$, Mix Intra : $4.7° \pm 0.1$, Son POM : $4.8° \pm 0.1$, tous les $p = 1$).

Regardons à présent si l'effet du son mesuré pour la catégorie Visages est modifié après un événement sonore saillant.

Après un événement sonore

Les modèles DESA et Energie présentés section 2.2.1 permettent d'extraire pour chacune de nos bandes-son une courbe de saillance sonore, dont les principaux pics repèrent les événements sonores dudit signal. Nous comparons les valeurs moyennes des métriques présentées ci-dessus avec celles comprises dans des fenêtres temporelles suivant les pics de saillance sonore de 5, 10 et 15 frames. N'ayant mesuré d'effet des conditions expérimentales que pour la catégorie Visages, nous ne présentons ici que les résultats de celle-ci. Par soucis de clarté et comme les résultats ne varient ni selon le modèle utilisé (DESA ou Energie) ni selon la taille de la fenêtre, nous ne présentons que ceux obtenus avec le DESA et une fenêtre de 10 frames. Pour chaque métrique, nous avons mené une ANOVA à deux facteurs intra (les valeurs de la métrique en moyenne et après les pics de saillance sonore ; la condition expérimentale).
Dans la catégorie Visages, la distance au centre augmente après les pics de saillance sonore (en moyenne : $6.8° \pm 0.2$, après pics : $6.9° \pm 0.2$, $F(1,14) = 10.1$, $p = .007$).

Il en va de même pour les durées de fixation (en moyenne : 379 ms ± 10, après pics : 410 ms ± 14, $F(1,71) = 38.8$, $p < .001$).
Pour les autres métriques nous n'avons pas mesuré d'effet des pics de saillance sonore.

La seule catégorie visuelle à présenter une interaction systématique avec son contenu sonore, à la fois en moyenne et après un événement sonore est la catégorie Visages. L'occurrence d'un évènement sonore saillant amène les participants à s'éloigner davantage du centre de l'écran, comme pour chercher et regarder une éventuelle source sonore. Ce comportement semble s'accorder avec les plus longues fixations enregistrées après les pics de saillance sonore : une fois la source sonore localisée, les observateurs la regardent plus longtemps afin de comprendre les causes ou de suivre les conséquences de l'événement l'ayant amené à produire un tel son. Toutefois, il est étrange de constater l'absence d'interaction significative entre les pics de saillance et les conditions expérimentales. On aurait pu penser que les pics de saillance sonore affecteraient plus volontiers le regard lorsque l'information sonore est liée à ce qui se passe à l'image, comme c'est le cas dans la condition Originale, ou, dans une moindre mesure, dans la condition Mix Intra.

Jusqu'à présent, nous avons étudié les différences induites par les catégories visuelles et conditions expérimentales sur certaines caractéristiques des mouvements oculaires. Cette approche ne permet pas de prendre en compte l'information portée par les stimuli, de comprendre quelles caractéristiques de la scène sollicitent les mouvements oculaires enregistrés. Les attributs visuels les plus fixés varient-ils selon la catégorie visuelle ? Selon la condition expérimentale ? Nous proposons de répondre à ces questions à travers le prisme de la modélisation statistique.

3.3 Choix de modèle par sélection de variables

Lors de l'exploration de scènes naturelles, les attributs bas niveau présents dans les régions fixées par les observateurs diffèrent significativement de ceux présents dans les régions non fixées. De nombreux outils ont été développés pour trouver les attributs attirant le mieux le regard des observateurs. Certains auteurs comparent les propriétés locales de la scène (contraste, luminosité, densité de contours...) entre des régions choisies aléatoirement et au voisinage des fixations, au moyen de statistiques d'ordre supérieures [Krieger *et al.* 2000, Parkhurst & Niebur 2003]. D'autres calculent des cartes représentant les distributions spatiales des différents attributs et les comparent avec des cartes de densité de positions oculaires *via* des métriques telles que l'aire sous une courbe Receiver Operating Characteristic (ROC) [Tatler *et al.* 2005, Mital *et al.* 2010], la divergence de Kullback-Leibler [Itti 2005], la différence d'histogrammes [Carmi & Itti 2006], le Normalized Scanpath Saliency [Peters

et al. 2005], voir [Le Meur & Baccino 2013] pour une revue complète.
Dans cette section nous abordons ce problème par le biais de la modélisation statistique, encore peu utilisée dans la littérature malgré ses nombreux avantages. Nous l'appliquons à nos données de manière à quantifier et comparer l'importance de différents attributs visuels bas niveau selon les différentes catégories visuelles et conditions expérimentales. Nous faisons par exemple l'hypothèse que les attributs dynamiques expliqueront mieux les fixations enregistrées dans la catégorie POM que dans la catégorie Paysages, où le mouvement est moins présent et où les attributs statiques pourraient prépondérer. Cette analyse pourrait également nous permettre de préciser l'effet des conditions expérimentales mesurées dans la catégorie Visages. Certains attributs bas niveau sont-ils davantage fixés dans la condition Originale que dans les autres conditions ? Ou les différences constatées reposent-elles sur des facteurs de plus haut niveau ?

3.3.1 Principes théoriques

Soit une variable quantitative $\mathbf{Y}$ pouvant être modélisée ou approchée par une combinaison linéaire de p variables quantitatives $\mathbf{X}_{i \in [1..p]}$ connues (les attributs), pondérées par des poids $\boldsymbol{\beta} = \beta_{i \in [1..p]}$.

$$Y = \sum_{i=1}^{p} \beta_i X_i \tag{3.1}$$

Le but de la modélisation statistique est l'estimation de la famille $\boldsymbol{\beta}$ expliquant au mieux la variable $\mathbf{Y}$. Ceci peut permettre de poursuivre trois objectifs :

Descriptif, si on veut rechercher de façon exploratoire la relation entre $\mathbf{Y}$ et les $\mathbf{X}_\mathrm{i}$, afin par exemple de sélectionner le sous-ensemble $\mathbf{X}_{j \in [1..q]}$ le plus pertinent, c'est-à-dire celui dont les $\beta_{j \in [1..q]}$ associés sont les plus grands (à la manière d'une Analyse en Composantes Principales).

Explicatif, si l'on a une connaissance *a priori* de la relation entre $\mathbf{Y}$ et les $\mathbf{X}_\mathrm{i}$, et que l'on veut confirmer, infirmer ou préciser cette dernière.

Prédictif, si l'on veut construire un modèle prédisant $\mathbf{Y}$ à partir d'un nombre plus ou moins restreint de variables $\mathbf{X}_\mathrm{i}$.

Dans le domaine d'étude qui nous intéresse, $\mathbf{Y}$ peut être la carte de densité des positions oculaires enregistrées sur une image donnée, et les $\mathbf{X}_\mathrm{i}$ peuvent être des cartes représentant des attributs connus pour attirer le regard (Figure 3.5). Estimer $\boldsymbol{\beta}$ permet alors de quantifier frame après frame l'importance de certains attributs par rapport à d'autres, et de la comparer selon les conditions expérimentales. Au cours de cette thèse, nous avons suivi deux approches différentes : l'optimisation de la vraisemblance (Espérance - Maximisation) et la régression pénalisée (Lasso).

Table 3.2 – Avantages de l'estimation par Espérance-Maximisation et par Lasso.

	Avantages
Espérance Maximisation	estimation des variables cachées
	coefficients entre 0 et 1
	utilisé dans la littérature
Lasso	parcimonieux
	pas d'initialisation
	peu coûteux

3.3.1.1 Espérance - Maximisation

L'algorithme d'Espérance - Maximisation (EM) est une procédure itérative, dont le but est de converger vers le maximum de vraisemblance des paramètres d'un mélange de variables [Dempster *et al.* 1977]. Cette méthode est très souvent utilisée pour le partitionnement automatique de données (e.g. mélange de gaussiennes) en *machine learning* ou en *computer vision*, et de récentes études l'ont appliquée avec succès à la modélisation de l'attention visuelle sur des scènes statiques [Vincent *et al.* 2009, Couronné *et al.* 2010, Ho-Phuoc *et al.* 2010, Gautier & Le Meur 2012, Queste-Devillez 2014]. Un grand avantage de cette méthode est qu'elle permet d'estimer aussi certaines variables cachées du problème, comme le centre ou la variance d'une gaussienne. A notre connaissance, l'EM n'avait jamais été utilisé avec des scènes dynamiques.
La procédure suivante est appliquée pour chaque frame. Tout d'abord, les cartes $\mathbf{X}_i$ et $\mathbf{Y}$ sont chacune converties en distribution de probabilité. Après une initialisation des paramètres $\boldsymbol{\beta}$, deux étapes se succèdent jusqu'à convergence :

Espérance : on fait l'hypothèse que le modèle courant (c'est-à-dire la combinaison des paramètres $\boldsymbol{\beta}$ courante) est correct, et nous calculons l'espérance de la vraisemblance (ici grâce aux positions oculaires de l'ensemble des participants).

Maximisation : on modifie $\boldsymbol{\beta}$ de manière à maximiser la quantité calculée à l'étape précédente.

Au final, nous disposons du poids des variables $\mathbf{X}_i$ pour chaque frame de chaque vidéo dans chaque condition expérimentale. Pour plus de détails, voir l'implémentation de l'algorithme, Annexe D.

3.3.1.2 Lasso

Le Least Absolute Shrinkage and Selection Operator (Lasso) [Tibshirnani 1996] est une méthode statistique largement répandue en génétique, pour déterminer sur quels gènes s'appuie l'expression d'un caractère donné [Yi & Xu 2008], ou en reconnaissance de forme, pour choisir les attributs caractérisant le mieux l'objet traité [Wright *et al.* 2010]. A part dans une belle étude de Baddeley et Tatler, cette approche n'a curieusement jamais été utilisée par la communauté vision [Baddeley & Tatler 2006].
Il s'agit d'une variante régularisée de la célèbre méthode des moindres carrés [Legendre 1805, Gauss 1809]. Rappelons que cette méthode propose, pour N observations, de choisir la famille $\boldsymbol{\beta}$ minimisant la somme des carrés des résidus :

$$\boldsymbol{\beta}^{MC} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (Y_i - \sum_{j=1}^{p} \beta_j X_{ij})^2 \right\} \tag{3.2}$$

La méthode des moindres carrés pose deux problèmes. Le premier est la précision de l'estimation : une régression par les moindres carrés présente souvent un faible biais dans l'estimation de l'espérance, mais une grande variance. En effet, lorsqu'une variable $\mathbf{X}_\mathrm{i}$ est corrélée à une autre variable $\mathbf{X}_\mathrm{j}$, les paramètres issus des moindres carrés β_i^{MC} et β_j^{MC} peuvent présenter une importante variance. Par exemple, la très grande valeur positive de l'un peut être compensée par la très grande valeur négative de l'autre. Le second problème est l'interprétabilité de la régression : cette méthode prend en compte l'ensemble des variables d'entrée, or il est souvent souhaitable de ne conserver que celles qui contribuent significativement à l'observation. Le Lasso propose, lors de la régression, de contraindre la somme des valeurs absolues des coefficients $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}^{Lasso} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (Y_i - \sum_{j=1}^{p} \beta_j X_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{3.3}$$

avec λ une constante de pénalisation. Une formulation équivalente est de dire que la somme des valeurs absolues ne doit pas dépasser une certaine constante :

$$\boldsymbol{\beta}^{Lasso} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (Y_i - \sum_{j=1}^{p} \beta_j X_{ij})^2 \right\} \text{ avec } \sum_{j=1}^{p} |\beta_j| \leq \lambda \tag{3.4}$$

Contraindre la valeur absolue des poids a pour effet d'imposer une certaine parcimonie aux coefficients $\boldsymbol{\beta}$, ce qui résout en partie les problèmes précédents. Il est à noter que le Lasso est sensible à l'échelle des variables utilisées, il convient donc de les normaliser. Le Lasso est très similaire à la régularisation Tikhonov, ou régression d'arête [Tikhonov 1943]. Dans cette méthode, la pénalisation L_1 (somme des valeurs absolues) est remplacée par la pénalisation L_2 (somme des carrés). L'avantage de

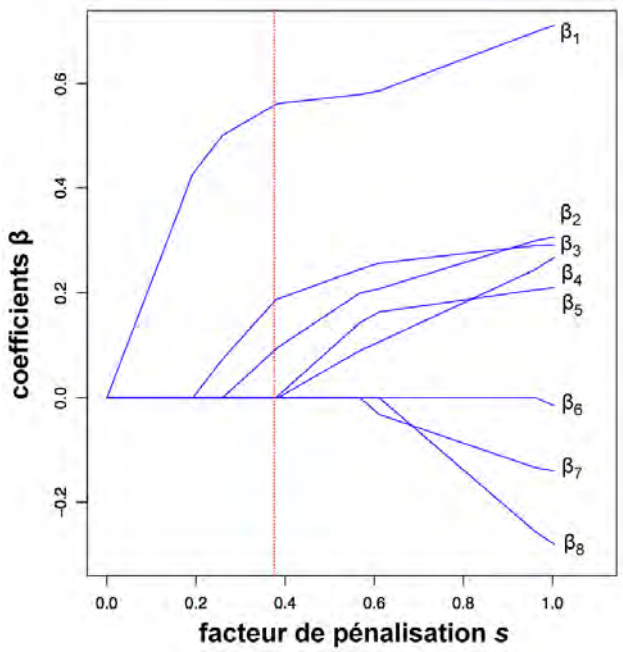
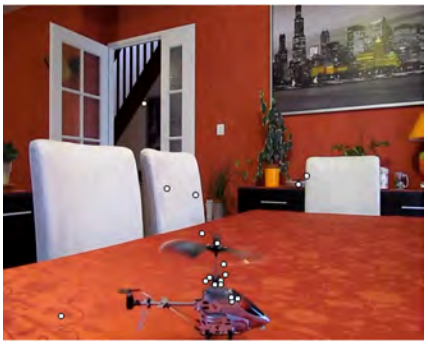


Figure 3.4 – Estimation par la méthode Lasso de 8 coefficients $\beta_{i\in[1..8]}$ en fonction du facteur de pénalisation $s = \lambda / \sum_{i=1}^{8} |\beta_i|$ (voir la formule 3.4). Un facteur $s = 1$ correspond à l'estimation des moindres carrés, $s = 0$ correspond à annuler tous les coefficients. La ligne verticale rouge correspond à la famille $\boldsymbol{\beta}$ dont le BIC est optimal : ici, seuls les paramètres β_1, β_2 et β_3 sont non nuls. Adapté de [Hastie *et al.* 2009], page 70.

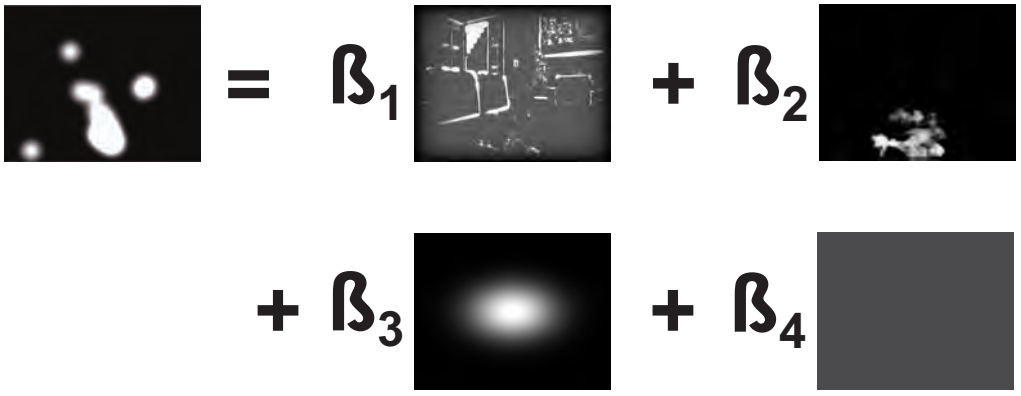
la pénalisation L_1 par rapport à la pénalisation L_2 est qu'elle est davantage parcimonieuse, c'est-à-dire qu'elle va davantage annuler les paramètres ne contribuant pas significativement aux observations [Ng 2004, Hastie *et al.* 2009]. Cette propriété fait du Lasso un précieux outil pour sélectionner des variables dans des modèles de grande dimension. Si $\lambda = 0$, on retrouve l'estimateur des moindres carrés. Si λ tend vers l'infini, on annule tous les β_i. La résolution du Lasso est un problème non linéaire, et de nombreux algorithmes efficaces ont été proposés. Dans notre étude, nous avons utilisé l'implémentation proposée dans la toolbox Sparse Statistical Modeling [Sjöstrand *et al.* 2012]. Il s'agit de faire varier λ de manière à pénaliser plus ou moins fortement les paramètres du problème. Ainsi, une famille $\boldsymbol{\beta}$ est déterminée pour chaque pas $\lambda + d\lambda$. Celle dont le modèle correspondant a le Critère d'Information Bayésien (BIC) le plus faible est choisie (ligne verticale rouge de la Figure 3.4). Le BIC est une mesure de la qualité d'un modèle statistique, prenant en compte à la fois la vraisemblance et le nombre de paramètres du modèle [Schwarz 1978]. Il s'écrit, pour un modèle M et une observation Y donnés,

$$BIC(M|Y) = -2 \log L(M|Y) + p \log n \tag{3.5}$$

avec $L(M|Y)$ la vraisemblance du modèle M compte tenu des observations Y (ici nous prenons la somme des carrés des résidus), p le nombre de paramètres du modèle



(a)



(b)

Figure 3.5 – **a -** Frame issue d'une vidéo appartenant à la catégorie visuelle UOM. Les points noirs et blancs correspondent aux fixations des participants ayant regardé cette vidéo dans la condition expérimentale Originale. **b -** Modélisation de la carte de densité des positions oculaires (à gauche) par une combinaison linéaire de la carte de saillance statique, de saillance dynamique, du biais de centralité, et d'une carte uniforme. Le but est ici d'estimer les poids β correspondant à chacune de ces cartes, en fonction de la catégorie sonore.

et n le nombre de points dans Y. Là aussi, le nombre de paramètres du modèle est pénalisé de manière à prévenir le sur-ajustement. Le modèle sélectionné sera donc celui dont le BIC est le plus faible. Il est à noter que les coefficients β du Lasso sont signés et que leur somme n'est pas forcément unitaire, ce qui rend leur interprétation délicate. Aussi, dans nos analyses, nous les avons normalisés entre 0 et 1.

3.3.2 Application à notre objet d'étude

Nous avons utilisé les méthodes statistiques précédemment évoquées afin de quantifier la capacité de différents attributs visuels à expliquer les positions oculaires enregistrées dans les différentes catégories visuelles et conditions expérimentales. Pour chaque frame de chaque stimulus nous avons généré cinq cartes différentes. La

carte $\mathbf{Y}$ que l'on cherche à modéliser est la carte de densité des positions oculaires définies au chapitre précédent, équation 1.2 (première carte de la Figure 3.5b). Les quatre cartes représentant les attributs $\mathbf{X}_i$ de notre modèle sont :

1. La carte de la voie statique du modèle de saillance visuelle proposé dans [Marat *et al.* 2009] et décrit section 1.2.3.2 (deuxième carte de la Figure 3.5b).
2. La carte de la voie dynamique du même modèle de saillance visuelle (troisième carte de la Figure 3.5b).
3. La carte représentant le biais de centralité. La plupart des études oculométriques ont constaté que les participants ont plus tendance à regarder le centre que la périphérie des stimuli. Différentes hypothèses ont été formulées à ce sujet, voir section 1.2.2.3. Comme proposé dans [Gautier & Le Meur 2012], nous avons modélisé le biais de centralité par une gaussienne bi-dimensionnelle centrée au milieu de l'écran : $N(0, \Sigma)$, avec $\Sigma = \left(\begin{smallmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{smallmatrix}\right)$ la matrice de covariance et σ_x^2, σ_y^2 la variance. Nous avons pris σ_x et σ_y proportionnels à la taille d'une frame (28°x 22.5°), et avons mené plusieurs analyses avec des valeurs s'échelonnant de $\sigma_x = 2°$ à $\sigma_x = 3.5°$ et de $\sigma_y = 1.6°$ à $\sigma_y = 2.8°$. La variation de ces valeurs ne modifiant pas significativement la modélisation, nous ne présenterons par la suite que les résultats obtenus avec $\sigma_x = 2.3°$ et $\sigma_y = 1.9°$ (quatrième carte de la Figure 3.5b).
4. Une carte uniforme, dans le cas où les trois cartes précédentes failliraient à expliquer une position oculaire (cinquième carte de la Figure 3.5b).

Résultats

Le but est donc d'estimer les coefficients $\boldsymbol{\beta}$ permettant de modéliser au mieux la carte de densité de positions oculaires. Cette estimation est réalisée via les algorithmes Lasso et EM. Comme précédemment, nous commençons par comparer les résultats obtenus entre les différentes catégories visuelles, indépendamment des conditions expérimentales (Figure 3.6). Nous avons mené une ANOVA sur les coefficients $\boldsymbol{\beta}$ à deux facteurs inter (les attributs du modèle (saillance statique, saillance dynamique, biais de centralité et carte uniforme) et les catégories visuelles (Visages, Paysages, UOM et POM)). Chaque niveau a 60 items (15 stimuli × 4 bandes-son).

Estimation Lasso

Il existe un effet principal des attributs ($F(3,177) = 272.9$, $p < .001$) : les quatre poids sont tous différents les uns des autres (tous les $p < .001$). L'effet de la catégorie visuelle n'est pas significatif ($F(3,177) = .82$, $p = .48$), contrairement à son interaction avec les attributs ($F(9,531) = 43$, $p < .001$).

Saillance statique : il n'y a pas de différences entre les catégories visuelles (tous les $p = 1$, sauf entre UOM et POM, $p = .44$).

Saillance dynamique : la catégorie Visages est supérieure à toutes les autres

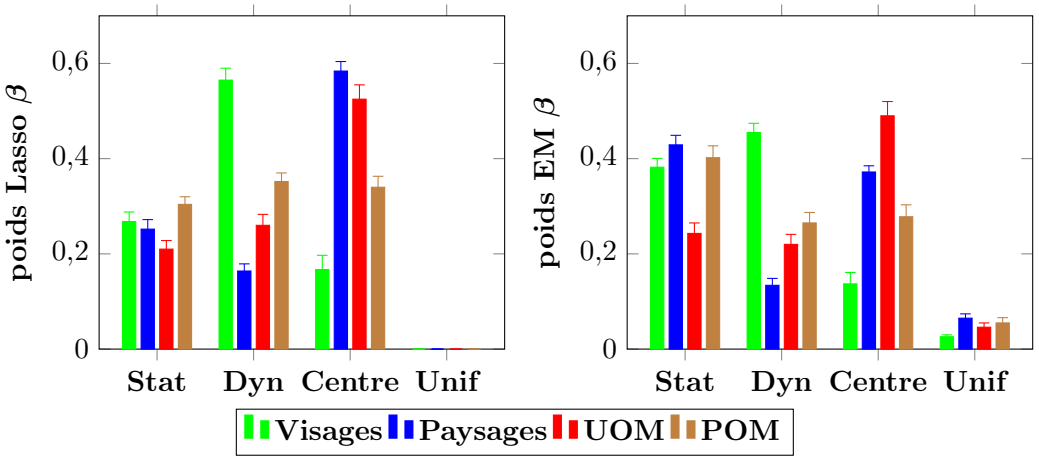


Figure 3.6 – Coefficients β obtenus par la méthode Lasso (à gauche) et Espérance-Maximisation (à droite) pour la saillance statique, la saillance dynamique, le biais de centralité et la carte uniforme dans chaque catégorie visuelle, toutes conditions expérimentales confondues. Les barres d'erreur correspondent aux erreurs standards.

(tous les $p < .001$), la catégorie Paysage n'est pas différente de la catégorie UOM ($p = .26$) mais est inférieure à la catégorie POM ($p < .001$). La catégorie UOM n'est pas différente de la catégorie POM ($p = .52$).
Biais de centralité : les catégories Paysages et UOM sont supérieures à toutes les autres (tous les $p < .001$), mais ne sont pas différentes l'une de l'autre ($p = 1$). Viennent ensuite la catégorie POM, puis la catégorie Visages dont le biais de centralité est inférieur à toutes les autres (tous les $p < .001$).
Carte uniforme : aucune différence entre les catégories visuelles (tous les $p = 1$).

L'évolution temporelle des coefficients de ces attributs sur les 100 premières frames de chaque stimuli est représentée Figure 3.7. Les coefficients de la carte uniforme étant nuls, nous ne les avons pas affichés. Comme pour la dispersion ou la distance au centre on note une latence d'environ 5 frames au début de chaque vidéo, correspondant au temps que mettent les observateurs à quitter le centre de l'écran pour commencer leur exploration. Durant cette période, l'attribut correspondant au biais de centralité est à son maximum ($\beta_{\text{centre}} = 1$) et les autres à leur minimum ($\beta_{\text{stat}} = \beta_{\text{dyn}} = 0$). Par la suite, les coefficients du biais de centralité et ceux de la saillance dynamique suivent une évolution inverse. La saillance dynamique (resp. le biais de centralité) augmente (resp. diminue) brutalement, pour ensuite se stabiliser autour d'une valeur moyenne, différente selon la catégorie visuelle. La catégorie Visages se distingue des autres par un poids moyen particulièrement élevé pour la saillance dynamique, et particulièrement bas pour le biais de centralité. Pour la saillance statique, une rapide augmentation est également visible entre les frames 5 et 20, mais la suite de l'évolution présente de fortes fluctuations.
Saillance statique : nous avons mené une ANOVA à un facteur inter : la catégorie

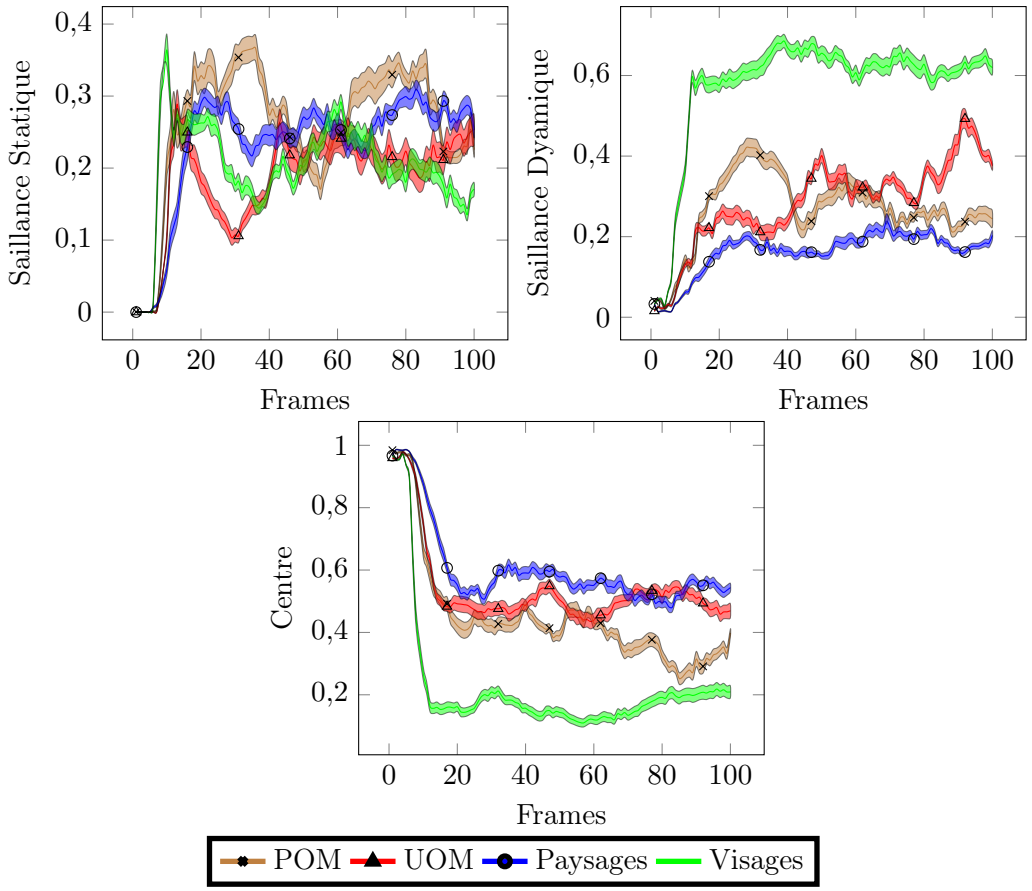


Figure 3.7 – Evolution temporelle des coefficients des attributs "saillance statique", "saillance sonore" et "biais de centralité" estimés avec la méthode Lasso. Les coefficients de la carte "uniforme" ne sont pas représentés car nuls. Les barres d'erreur correspondent aux intervalles de confiance à 95%

visuelle, et un facteur intra : le temps (4 niveaux : phase 1 des frames 1 à 5, phase 2 des frames 5 à 25, phase 3 des frames 25 à 50, et phase 4 de la frame 50 à la fin des vidéos). Chaque niveau a 15 items : le poids de la saillance statique de chaque vidéo moyenné sur les conditions expérimentales. Il existe un effet principal de la catégorie visuelle ($F(3,42) = 5.5$, $p = .003$), du temps ($F(3,42) = 461.6$, $p < .001$) et de l'interaction ($F(9,126) = 4.9$, $p < .001$). Pour toutes les catégories visuelles, le poids de la saillance statique augmente de la phase 1 à la phase 2 (tous les $p < .001$) mais se stabilise par la suite (pas de différence significative entre les phases 2 et 4). **Saillance dynamique :** nous avons mené la même ANOVA à un facteur inter : la catégorie visuelle, et un facteur intra : le temps. Il existe un effet principal de la catégorie visuelle ($F(3,42) = 56.2$, $p < .001$), du temps ($F(3,42) = 340.5$, $p < .001$) et de l'interaction ($F(9,126) = 29.7$, $p < .001$). Pour toutes les catégories visuelles, le poids de la saillance dynamique augmente de la phase 1 à la phase 2 (tous les $p < .001$) mais se stabilise par la suite (pas de différence significative entre les phases 2 et 4).

Biais de centralité : nous avons mené la même ANOVA à un facteur inter : la catégorie visuelle, et un facteur intra : le temps. Il existe un effet principal de la catégorie visuelle ($F(3,42) = 52.5$, $p < .001$), du temps ($F(3,42) = 749.9$, $p < .001$) et de l'interaction ($F(9,126) = 20.4$, $p < .001$). Pour toutes les catégories visuelles, le poids du biais de centralité diminue de la phase 1 à la phase 3 (tous les $p < .05$) mais se stabilise par la suite (pas de différence significative entre les phases 3 et 4, tous les $p = 1$).

Estimation Espérance - Maximisation

Il existe un effet principal des attributs ($F(3,177) = 148.5$, $p < .001$) : les quatre niveaux sont tous différents les uns des autres (tous les $p < .001$, sauf la saillance statique et le biais de centralité, $p = 0.04$, et la saillance dynamique et le biais de centralité, $p = 0.012$). L'effet de la catégorie visuelle n'est pas significatif ($F(3,177) = 1.22$, $p = .3$), contrairement à son interaction avec les attributs ($F(9,531) = 33$, $p < .001$).

Saillance statique : la catégorie UOM est inférieure à toutes les autres (tous les $p < .001$), mais il n'y a pas de différences entre les autre catégories visuelles (tous les $p = 1$).

Saillance dynamique : la catégorie Visages est supérieure à toutes les autres (tous les $p < .001$), la catégorie Paysage n'est pas différente de la catégorie UOM ($p = .73$) mais est inférieure à la catégorie POM ($p = .004$). La catégorie UOM n'est pas différente de la catégorie POM ($p = 1$).

Biais de centralité : la catégorie UOM est supérieure à toutes les autres (tous les $p < .001$, sauf avec la catégorie Paysages, $p = .018$), suivie des catégories Paysages et POM qui ne sont pas significativement différentes l'une de l'autre ($p = .33$). La catégorie Visages a le biais de centralité le plus faible (tous les $p < .001$).

Carte uniforme : aucune différence entre les catégories visuelles (tous les $p = 1$).

L'évolution temporelle des coefficients estimés par la méthode Espérance - Maximisation n'est pas représentée ici, mais elle a la même allure que celle des coefficients estimés par la méthode Lasso.

Selon l'association audiovisuelle

Pour chaque catégorie visuelle, nous avons également mené une ANOVA à deux facteurs intra : les attributs du modèle et la condition expérimentale. Chaque niveau a 15 items (un par stimulus). Pour l'estimation Lasso comme pour l'estimation EM, et conformément aux résultats ci-dessus, chaque ANOVA a mis en évidence l'effet principal des attributs. Par contre aucun effet des conditions expérimentales ni de l'interaction n'a été constaté. Le détail des statistiques est disponible en Annexe C.2.

Afin d'étudier l'effet des pics de saillance sonore sur l'importance relative des attributs visuels, nous avons mené pour chaque catégorie visuelle une ANOVA à trois facteurs intra (la condition expérimentale ; les attributs du modèle ; les attributs en moyenne sur chaque stimulus, et après les pics de saillance sonore). Chaque niveau a 15 items (un par stimulus). Là aussi, conformément aux résultats précédents, chaque ANOVA a mis en évidence l'effet principal des attributs. Par contre aucun effet ni des conditions expérimentales ni des pics de saillance, ni des différentes interactions n'a été constaté.

3.4 Discussion

Au chapitre précédent, nous avions enregistré les mouvements oculaires de participants regardant des vidéos avec leurs bandes-son (condition AudioVisuelle) ou sans aucun son (condition Visuelle), et avions identifié plusieurs différences entre les deux conditions expérimentales. Nous avions également comparé les mouvements oculaires effectués au voisinage d'événements sonores particulièrement saillants à ceux effectués en moyenne tout au long des stimuli, mais n'avions trouvé aucune différence significative. Nous avions alors formulé l'hypothèse que le contenu audiovisuel de nos vidéos était trop disparate, et qu'il fallait mieux le contrôler : l'intégration audiovisuelle pourrait par exemple s'exprimer différemment dans une scène de paysage et dans une scène de conversation.
Ceci nous a conduit à l'expérience 2, présentée dans ce chapitre, pour laquelle nous avons construit une base de stimuli classés selon leur contenu audiovisuel. Dans cette section, nous discutons des résultats de cette expérience. Dans un premier temps, nous proposons une caractérisation des stratégies d'exploration en fonction du contenu visuel des stimuli, sans tenir compte de l'information sonore. Puis, nous nous penchons sur l'influence des conditions expérimentales.

3.4.1 Différents contenus visuels induisent différentes explorations

Depuis longtemps, nous savons que le système visuel a la capacité de classifier extrêmement rapidement une scène naturelle. Il est capable de détecter la présence dans une image de certaines catégories d'objets (animal, véhicule) dans un laps de temps inférieur à la durée d'une fixation [Potter 1976, Thorpe *et al.* 1996]. Cependant, bien peu d'études ont cherché à comprendre comment différentes catégories de contenu visuel pouvaient influer sur les stratégies d'exploration développées par les observateurs.

Table 3.3 – Relations constatées dans la littérature entre les valeurs de dispersion, de distance au centre (DtC), de durée de fixation (Durée Fix) et d'amplitude de saccade (Amp Sac), et la complexité de la scène explorée. Ici, la complexité d'une scène mesure la quantité d'information disponible. Une scène complexe (Visages, POM) présente de nombreuses et informatives régions d'intérêt (ROI), alors qu'une scène simple (Paysages, UOM) en contient peu ou pas du tout. (1) [Dorr *et al.* 2010] - (2) [Jansen *et al.* 2009] - (3) [Judd *et al.* 2011] - (4) [Mital *et al.* 2010] - (5) [Smith & Mital 2013] - (6) [Smith 2013] - (7) [Tseng *et al.* 2009] - (8) [Unema *et al.* 2005] - (9) [Mancas & Le Meur 2013]

	Augmente si complexe	**Diminue si complexe**
Dispersion	plus de compétition entre ROI$^{(3)}$	les ROI focalisent l'attention$^{(1,4,5,9)}$
DtC	regard s'éloigne du centre vers les régions d'intérêt$^{(5,4)}$	centre = stratégique pour capter un maximum de ROI$^{(1,6)}$
Durée Fix	les ROI attirent le regard plus longtemps$^{(1,5,9)}$	de nombreuses ROI à explorer$^{(2,8)}$
Amp Sac	si ROI dispersées$^{(1,5)}$	si ROI regroupées$^{(2,8)}$

3.4.1.1 Complexité de la scène

Dans ce chapitre, nous avons construit quatre catégories au contenu visuel et à la complexité variés. Par complexité, nous entendons la quantité d'information contenue dans une scène susceptible d'attirer l'attention des observateurs. La catégorie Paysages est la moins complexe, ses vidéos ne présentent pas de régions particulièrement saillantes et ne contiennent aucune information sémantique. Vient ensuite la catégorie Un Objet en Mouvement (UOM) dont les vidéos présentent une scène encore simple mais contenant une zone de forte saillance, celle correspondant à l'objet en mouvement. Les vidéos de la catégorie Plusieurs Objets en Mouvement (POM) sont plus complexes, les observateurs pouvant porter leur attention sur plusieurs objets d'intérêt simultanément en compétition. Enfin, la catégorie Visages est particulièrement riche en information dans la mesure où la perception des visages remplit une fonction sociale toute particulière (voir section 4.1).

Les différences que nous avons obtenues entre ces catégories peuvent être interprétées à la lumière de cette gradation de quantité d'information et de complexité. En effet, nous obtenons des durées de fixation plus longues pour les catégories les moins riches (Paysages et UOM), et plus courtes pour les plus complexes (Visages et POM). Nous pouvons formuler l'hypothèse selon laquelle plus il y a d'information dans une scène, plus cette dernière est explorée rapidement pour traiter toutes les ROI présentes.
Cette idée est soutenue par les résultats présentés dans [Unema *et al.* 2005], où

les durées de fixation et les amplitudes de saccade décroissent lorsque le nombre d'objets présents (et donc la complexité) augmente, ou encore dans [Jansen *et al.* 2009], où les auteurs constatent une diminution de la durée des fixations et de l'amplitude des saccades avec l'introduction d'information de disparité binoculaire. C'est également le cas dans [Mannan *et al.* 1995], où les durées de fixation croissent lorsque l'image est filtrée passe-bas (et donc lorsque l'information disponible diminue), mais ce résultat n'est pas forcément comparable aux autres dans la mesure où le statut de "scène naturelle" d'une image floue est contestable.
Nos résultats de dispersion et de distance au centre, étroitement liés, pourraient aussi être interprétés par ce prisme. Nous obtenons une dispersion maximale et une faible distance au centre (ou un fort biais de centralité) dans la catégorie Paysages : aucun objet n'attirant le regard plus qu'un autre, les observateurs ne développent pas de séquence de fixations similaires, et ne s'éloignent pas non plus du centre, lieu d'observation stratégique pour capter l'éventuel apparition d'un objet digne d'intérêt [Tatler 2007]. A l'inverse, la catégorie Visages présente le biais de centralité le plus faible, suivie de la catégorie POM, signe que les observateurs ont activement exploré ces stimuli, s'éloignant du centre de l'écran pour aller ensemble fixer les visages, porteurs d'une riche information sociale et sémantique [Yarbus 1967], ou les objets en mouvement, également grands attracteurs d'attention [Mital *et al.* 2010]. Ces résultats sur la dispersion sont en accord avec une étude portant sur la mémorabilité d'images statiques [Mancas & Le Meur 2013]. Bien qu'il s'agisse de deux caractéristiques différentes, il est possible de lier la complexité et la mémorabilité d'une scène [Isola *et al.* 2011]. Une image sera d'autant moins mémorable qu'elle présentera une distribution spatiale de saillance uniforme, sans objet particulièrement saillant, comme c'est le cas pour la catégorie visuelle Paysages.
D'autres études montrent que la dispersion entre les positions oculaires effectuées sur une scène dynamique est inférieure à celles effectuées sur une frame statique issue de la même scène, et présentant donc un contenu visuel similaire [Dorr *et al.* 2010, Smith & Mital 2013]. Si l'on considère qu'une scène dynamique est plus complexe qu'une scène statique (où l'information de mouvement a été supprimée), les résultats de ces papiers sont en accord avec les nôtres. A l'inverse, ces auteurs ont mesuré de plus petites saccades et de plus courtes fixations sur les scènes statiques que sur les scènes dynamiques : dans ces dernières, les observateurs seraient attirés par un nombre plus restreint de régions d'intérêt, et les fixeraient plus longtemps.
Cette contradiction avec nos résultats montre qu'il est hasardeux de généraliser trop vite les stratégies d'exploration constatées sur quelques scènes. Par exemple, il ne faut pas trop hâtivement conclure que plus il y a d'information dans une scène, mieux le regard est guidé et donc plus la dispersion est faible et la distance au centre importante. En effet, dans [Judd *et al.* 2011], les auteurs ont présenté à leurs participants des scènes statiques de différentes complexités et ont constaté que la dispersion augmente avec cette dernière. Leur interprétation est que pour les images "complexes", les nombreux détails favorisent la diversification des stratégies d'exploration, alors que pour les images "faciles", les observateurs vont tous fixer l'objet principal, car il n'y a souvent rien d'autre à regarder. C'est exactement ce

qui se passe pour notre catégorie UOM qui présente la dispersion et la distance au centre la plus faible : tous les regards se portent sur l'unique objet en mouvement, souvent situé au centre de l'image. Ce résultat est en contradiction avec celui d'une étude montrant que le biais de centralité ne dépend pas de la distribution spatiale des objets dans l'image [Tatler 2007]. Cependant, les objets d'intérêt des vidéos de la catégorie UOM étaient en mouvement, et donc plus à même d'attirer le regard vers eux que les objets statiques utilisés par Tatler *et al.*
Ce constat amène à considérer un autre facteur souvent présenté comme déterminant pour la stratégie d'exploration adoptée par les observateurs : la dynamique temporelle du stimulus.

3.4.1.2 Dynamique temporelle

Les travaux de certains auteurs montrent que les stratégies d'exploration varient considérablement avec le temps. Il existerait deux phases d'exploration visuelle : une phase de découverte de la scène (courtes fixations et grandes saccades), suivie d'une phase d'exploration plus détaillée (longues fixations et petites saccades) [Buswell 1935, Antes 1974, Unema *et al.* 2005, Velichkovsky *et al.* 2005, Pannasch *et al.* 2008, Mills *et al.* 2011]. Dans [Follet *et al.* 2011], les auteurs nuancent ces résultats. Ils comparent les amplitudes de saccade et les durées de fixation de participants regardant des scènes naturelles statiques classées selon quatre catégories : rues, bords de mer, montagnes et paysages ouverts. Grâce à une méthode de classification (*k-means*), ils mettent en évidence deux populations de saccades aux amplitudes moyennes bien distinctes (2.5° et 10.5°). En étudiant l'évolution temporelle de la probabilité d'occurrence de ces deux groupes, Follet *et al.* constatent qu'ils ne sont pas séquentiels, mais coexistent tout au long de l'exploration. De plus, aucune différence n'est mesurée entre les différentes catégories visuelles.
Dans cette étude, nous appuyons la nuance de ces deux phases d'exploration, et montrons qu'elles dépendent notamment du contenu visuel. Pour les catégories Visages et Paysages, nous avons constaté une légère diminution des amplitudes de saccade avec le temps, alors que pour les catégories UOM et POM, ces dernières se stabilisent dès le $5^{\text{ème}}$ rang de saccade, soit après le biais induit par la croix de fixation initiale. Quelle que soit la catégorie visuelle, les durées de fixation se stabilisent également dès le début de l'exploration.
Une explication à cette contradiction réside peut être dans la nature des stimuli utilisés. Alors que les scènes utilisées par l'ensemble de ces auteurs étaient statiques, les nôtres sont dynamiques. Comme nous l'avons montré dans le chapitre précédent, les explorations de ces deux types de scène sont bien différentes. La présence de changements de plan et le constant renouvellement de l'information présente à l'image modifient drastiquement l'exploration, bridant la diversification des stratégies d'exploration et induisant une stabilisation des différents paramètres des mouvements oculaires.

A notre connaissance, la seule étude à s'être intéressée l'évolution temporelle de ces paramètres sur des scènes dynamiques est celle présentée dans [Smith & Mital 2013]. En analysant les mouvements oculaires effectués sur des vidéos de personnes évoluant dans divers contextes naturels, les auteurs ont constaté une diminution des amplitudes de saccade durant les deux premières secondes d'exploration, puis une stabilisation de ces dernières. Ils ont également mesuré une rapide augmentation de la durée des fixations durant les deux premières secondes, puis une augmentation moins franche mais toujours significative durant la fin de l'exploration. Mise à part cette dernière légère augmentation des durées de fixation, ces résultats sont semblables à ceux que nous obtenons pour la catégorie Visages, dont le contenu visuel est le plus proche de celui des stimuli utilisés par Smith *et al.*

L'évolution temporelle de la capacité de la saillance bas niveau à guider le regard a également fait l'objet de nombreux débats dans la communauté. Certains auteurs ont constaté une baisse de la saillance des zones fixées avec le temps, interprétant cela comme une diminution de l'influence des facteurs bas niveau sur l'allocation attentionnelle [Itti & Koch 2000, Parkhurst *et al.* 2002, Carmi & Itti 2006, Marat *et al.* 2009]. A part Parkhurst *et al.*, aucun de ces auteurs n'a testé statistiquement cette supposée décroissance temporelle. De plus, la méthodologie développée par ces derniers a fait l'objet de vives critiques [Tatler *et al.* 2005], et plusieurs équipes ont depuis au contraire mis en évidence une stabilisation de la saillance des zones fixées avec le temps [Jansen *et al.* 2009, Wang *et al.* 2010]. Nos travaux s'inscrivent dans la lignée de ces dernières études. Pour toutes les catégories visuelles, les poids de la saillance statique comme de la saillance dynamique augmentent rapidement au tout début de l'exploration, puis se stabilisent autour d'une valeur moyenne. Cependant, nos résultats indiquent qu'à part pour la catégorie Visages, les attributs de saillance bas niveau prédisent peu efficacement les positions oculaires, ces derniers étant à un niveau comparable à celui du biais de centralité. Après la mise en place de l'exploration, cette contre-performance est temporellement stable.

En résumé, les paramètres des mouvements oculaires et leurs évolutions temporelles sont très largement influencés par la catégorie visuelle explorée, ainsi que par la nature du stimulus (statique ou dynamique). De nombreux facteurs étant impliqués, il est difficile de faire émerger une tendance générale, comme l'indique la Table 3.3. Il apparait donc crucial de bien contrôler le contenu visuel des stimuli utilisés lorsque l'on teste l'effet d'une variable sur les paramètres des mouvements oculaires, et de ne pas présenter comme une vérité universelle les résultats obtenus avec un certain type de scène. Après avoir discuté des différentes stratégies d'exploration induites par le contenu visuel des stimuli, intéressons-nous à l'influence des conditions expérimentales.

3.4.2 Contenu sonore

Puisque la façon d'explorer une scène dépend tant de son contenu visuel, nous pouvons penser que le contenu sonore joue lui aussi un rôle déterminant dans la perception audiovisuelle. Paradoxalement, nous n'avons trouvé que peu de différences entre les conditions expérimentales, tant en moyenne sur l'ensemble des vidéos qu'après des événements auditifs particulièrement saillants. Ces résultats laissent à penser que dans les stimuli audiovisuels avec lesquels nous avons travaillé, le regard est principalement guidé par les attributs visuels de la scène, et que la modalité sonore ne joue qu'un rôle marginal.
Cependant, une catégorie visuelle se distingue des autres : les Visages. Dans cette catégorie, les observateurs semblent adopter une exploration particulièrement active, comme en témoigne la grande distance au centre. Mais surtout, cette catégorie est la seule dans laquelle existe un effet clair des conditions expérimentales : dans la condition Originale, la dispersion entre les positions oculaires est plus faible et les amplitudes de saccade plus courtes que dans les autres conditions. Ce résultat est cohérent avec celui d'une autre étude étudiant les mouvements oculaires de participants visionnant des vidéos avec leur bande-son originale ou sans aucun son [Song *et al.* 2013]. Parmi un jeu de bande-son très varié (parole, chant, cris d'animaux, véhicules, musique, explosion...), la différence la plus importante entre deux conditions expérimentales (avec ou sans son) a été mesurée pour les vidéos dont le contenu sonore comportait une voix humaine.
Mais d'où vient cette spécificité des visages ? La modélisation statistique que nous avons utilisée montre que le poids de la saillance dynamique est bien plus élevé dans cette catégorie que dans les autres, ce qui indique que les objets en mouvement ont spécifiquement attiré les regards. Cependant nous n'avons pas constaté d'effet des conditions expérimentales sur les poids des attributs visuels bas niveau. Ceci suggère que les différences constatées entre les mouvements oculaires enregistrés dans la condition Originale et ceux enregistrés dans les autres conditions expérimentales sont portées par des facteurs de plus haut niveau. Pour y voir plus clair, nous devons approfondir notre connaissance de la littérature de la perception audiovisuelle de la parole. Dans le chapitre suivant, nous allons donner un aperçu de cette immense littérature, et extraire de nos données de nouveaux indices permettant d'expliquer l'effet des conditions expérimentales sur les stimuli de la catégorie Visages.

Chapitre 4

Les visages, des objets audiovisuels particuliers

Oui, c'est Agamemnon, c'est ton roi qui t'éveille ;
Viens, reconnais la voix qui frappe ton oreille.

Jean Racine, *Iphigénie en Aulide* (1674)

Sommaire

Ici, nous développons les résultats de la catégorie Visages de l'expérience 2 présentée au chapitre précédent. Afin de mieux les comprendre, nous apportons un bref état de l'art sur la perception audiovisuelle de la parole. Puis, nous comparons nos résultats avec ceux présentés dans la littérature, et proposons des métriques originales spécifiquement adaptées à la compréhension des stratégies d'exploration des visages. Enfin, nous proposons une modélisation statistique de ces dernières similaire à celle présentée section 3.3.2, mais incluant les visages parmi les attributs visuels.

4.1 Etat de l'art sur la perception et l'exploration des visages

Nous nous intéressons ici à la perception audiovisuelle de la parole *via* un médium bien précis : le visage du locuteur. Dans un premier temps, nous rappelons les principaux mécanismes à l'œuvre lors de la perception et de l'exploration de visages silencieux. Puis, nous entrons dans le vif du sujet en exposant les principaux résultats issus des études portant sur l'intégration audiovisuelle de la parole ainsi que sur les mouvements oculaires qu'elle provoque ou sur lesquels elle s'appuie.

4.1.1 Perception et exploration de visages silencieux

Les visages jouent un rôle social très particulier. Leurs bonnes détection, reconnaissance, et interprétation sont fondamentales dans la vie en société. L'étude de leur perception est aussi vieille que l'oculométrie elle-même : dès les années trente Buswell met en évidence la propension naturelle que nous avons à fixer les visages en général, et les yeux en particulier [Buswell 1935, Yarbus 1967]. Ce phénomène a depuis été répliqué de nombreuses fois [Haaf & Bell 1967, Birmingham *et al.* 2009], et des études en neuroimagerie ont établi l'existence de plusieurs aires cérébrales dédiées au traitement des visages (par exemple la *Fusiform Face Area*), [Kanwisher *et al.* 1997,Haxby *et al.* 2000]. Qu'est-ce qui rend les visages si particuliers ? Leur perception est-elle guidée par des attributs liés à leur structure (spectre, contraste...), ou à un traitement plus haut niveau ?

4.1.1.1 Perception

Dans un paradigme de recherche visuelle (une cible au milieu de distracteurs), deux types de comportements sont observés. Lorsque le temps de réaction est indépendant du nombre de distracteurs, on dit que la cible est détectée en "parallèle" *via* un mécanisme pré-attentif : il y a un effet *pop-out* [Treisman & Gelade 1980, Wolfe 1994]. Un tel mécanisme s'observe généralement lorsque la cible diffère significativement de ses distracteurs (une cible rouge parmi des distracteurs verts par exemple). Lorsque le temps de réaction augmente linéairement avec le nombre de distracteurs, la recherche visuelle est dite "séquentielle". Ce type de recherche opère lorsque les attributs élémentaires (couleur, orientation...) de la cible et des distracteurs sont semblables. Selon la conception classique de la structure du cortex visuel, ces attributs élémentaires sont détectés dans ses premières aires (V1, V2, [Hubel & Wiesel 1959]). Les attributs détectés en parallèle lors d'un paradigme de recherche visuelle ont ainsi été associés à un traitement pré-attentif, bas niveau. Cependant, d'autres études remettent cette approche en question. Par exemple, certains auteurs ont montré que des attributs plus complexes (et donc plus haut

niveau) tels ceux liés à la vision 3D produisent également un effet *pop-out*, ce qui invaliderait l'équivalence entre recherche en parallèle et traitement bas niveau [Enns & Rensink 1991]. Un autre exemple de traitement haut niveau en parallèle est justement la perception des visages. Crouzet *et al.* ont demandé à des participants de fixer le centre d'un écran en les informant que deux images allaient apparaitre, une contenant un visage, l'autre un véhicule. Les sujets avaient pour tâche de fixer le plus rapidement possible celle contenant le visage, ce qui ne leur a pas demandé plus de 100-110 ms [Crouzet *et al.* 2010]. Cette latence est bien plus courte que celles mesurées dans des tâches *"go / no go"* similaires impliquant par exemple des animaux *versus* des moyens de transport (au moins 250 ms, [VanRullen & Thorpe 2001]). Ceci indique l'existence d'une voie de traitement *bottom-up* rapide spécifique aux visages.
Par ailleurs, Hershler & Hochstein ont montré qu'une cible appartenant grossièrement au concept de "visage" (face dessinée, ou dont les parties intérieures ou extérieures ont été floutées...) présentée au sein de distracteurs très différents les uns des autres (et donc dont aucun attribut élémentaire n'a pu être utilisé pour distinguer la cible des distracteurs) était traitée en parallèle. Cette capacité à généraliser la notion de "visage" à de nombreux types de représentation cachés parmi de nombreux types de distracteurs reflète bien un traitement perceptif haut niveau [Hershler & Hochstein 2005]. Il a également été montré qu'il est difficile, même lorsque nous le demandons explicitement, d'inhiber ce dernier. Lorsqu'un visage est présent dans une image, nous y dirigeons notre attention visuelle de manière réflexe au détriment de tout autre objet, au moins durant les premiers instants après l'apparition de la scène [Bindemann *et al.* 2007].
Pour résumer, la perception des visages au sein d'une scène semble être guidée de manière *bottom-up* directement vers des aires corticales haut niveau spécialisées, permettant une détection parallèle efficace.
Une fois un visage détecté et localisé, l'analyse des informations qu'il véhicule peut commencer. Comment explorons-nous les visages ?

4.1.1.2 Exploration

Même les plus fervents adeptes des moyens de communication numériques en conviendront : rien n'est plus efficace pour transmettre un message qu'une conversation en face à face, "entre quatre yeux". En effet, durant ce type d'échange, le visage des interlocuteurs transmet un certain nombre d'informations supplémentaires ou modulant celles contenues dans le discours, tant au niveau social et émotionnel que linguistique [Richardson *et al.* 2009, Bailly *et al.* 2010, Bailly *et al.* 2012]. Pour percevoir toutes ces informations, nous explorons continuellement le visage de nos interlocuteurs. Cette section s'attache à décrire ces stratégies d'exploration.
Comme nous l'avons dit en introduction, nous savons depuis les premières études oculométriques que les yeux sont de puissants attracteurs de l'attention (voir [Birmingham *et al.* 2009] pour un état de l'art). Quelle est l'origine de cette attirance

particulière vers les yeux ? Est-elle liée à la saillance visuelle, à une fonction sociale, à une stratégie d'acquisition de l'information exprimée par le visage ? La première hypothèse a été facilement réfutée. Birmingham *et al.* ont présenté des images statiques de visages dans des scènes complexes et ont montré que les premières fixations étaient dirigées vers les visages (et particulièrement les yeux) alors même que ces derniers n'étaient pas saillants (au sens de [Itti & Koch 2000], voir [Birmingham & Kingstone 2009]). De plus, les yeux étaient d'autant plus fixés que les personnes dans la scène étaient nombreuses, ce qui suggère que ces derniers attirent non pas pour leur saillance visuelle mais bien en raison de l'information sociale qu'ils véhiculent [Foulsham *et al.* 2010]. Cependant, il a été montré que les yeux ne sont pas toujours fixés en priorité, et que le déploiement de l'attention sur le visage dépend grandement de l'intention de l'observateur (ou de sa tâche) [Itier *et al.* 2007,Buchan *et al.* 2007,Hui-wen Hsiao & Cottrell 2008,Eisenbarth & Alpers 2011,Võ *et al.* 2012]. En effet, de manière assez logique, les zones fixées dépendent du type d'information recherchée : nous n'adoptons pas la même stratégie d'exploration si nous essayons d'identifier une personne, de comprendre ce qu'elle dit ou de déterminer son état émotionnel. Pour quantifier ces différentes stratégies, Gosselin & Schyns ont introduit la technique dite des "*Bubbles*", reprise par la suite par de nombreux auteurs [Gosselin & Schyns 2001]. Cette technique mesure les performances de sujets pour une tâche donnée avec des visages dont certaines zones ont été floutées ou masquées. Les régions provoquant les plus grosses chutes de performance sont alors considérées comme cruciales pour la tâche considérée. Ce paradigme a par exemple été utilisé pour montrer que l'information au voisinage de la bouche sert à identifier la joie, alors que la peur est davantage exprimée par les yeux [Smith *et al.* 2005]. Une étude très récente a également mis en évidence une forte variabilité dans l'exploration des visages selon les individus [Mehoudar *et al.* 2014]. Selon ces auteurs, la façon d'explorer un visage serait propre à chacun, et constituerait un véritable "trait comportemental", très stable dans le temps.
Il n'existe donc pas de carte universelle de l'exploration des visages, cette dernière étant à la fois idiosyncratique et dépendante des circonstances. Par la suite, nous nous plaçons dans un contexte de perception audiovisuelle de la parole : comment perception de la parole et exploration des visages interagissent-elles ?

4.1.2 Perception audiovisuelle de la parole

Imaginez-vous à un pot de thèse en train d'essayer de suivre les commentaires de votre collègue, au milieu de dizaines d'autres conversations et du bruit de la construction du tramway s'infiltrant par la fenêtre. En premier lieu, on peut penser que nous discriminons du bruit ambiant le discours sur lequel nous portons notre attention sur la base d'une analyse de la scène auditive seule, en séparant les nombreux signaux en objets sonores distincts [Bregman 1990]. Mais de nombreuses études ont montré que nous nous aidons aussi d'indices appartenant à d'autres modalités, et notamment à la modalité visuelle [Sumby & Pollack 1954,Ross *et al.* 2007].

Ventriloquie

Le ventriloquisme est une ancienne pratique religieuse des oracles grecs. Appelée gastromancie, les sons produits par l'estomac étaient interprétés comme des voix d'outre-tombe produites par des morts ayant élu domicile dans le ventre du ventriloque. Cette illusion a été reprise par de nombreux comédiens (ci-contre : Terry Bennett), ou simplement lorsque nous regardons un film : les voix semblent provenir de la bouche des protagonistes plutôt que des véritables haut-parleurs [Altman 1980].

Mais comment quantifier l'interaction audiovisuelle dans la perception de la parole ?

Cette dernière a été étudiée selon deux principaux paradigmes. Le premier évalue l'importance de l'interaction audiovisuelle en comparant le comportement ou l'activité cérébrale de personnes face à un stimulus bimodal, *versus* unimodal, ou dont une des modalités est altérée (bruitée, floutée). Le second paradigme rend compte de la compétition existant entre les deux modalités en utilisant des stimuli bimodaux congruents ou incongruents (l'image correspond ou non au signal sonore perçu). Ces paradigmes ont fourni un cadre expérimental à de nombreuses études abordant de manières très diverses la perception audiovisuelle de la parole, et dont nous allons évoquer les principaux axes.

4.1.2.1 Intégration

Nous sommes capable d'intégrer un signal de parole avec l'information visuelle portée par le visage associé : voir notre locuteur parler améliore considérablement la perception de son discours. Mais quelles sont les conditions nécessaires à cette intégration audiovisuelle ? Comme il a été rappelé dans l'état de l'art (section 1.4), l'intégration multimodale est souvent théorisée dans le cadre de l'"hypothèse d'unité". Cette hypothèse stipule que plus les caractéristiques (par exemple spatio-temporelles) de stimuli issus de différentes modalités sont liées, plus il est probable que le cerveau les interprète comme provenant d'une source commune [Welch & Warren 1980, Bertelson & de Gelder 2004, Calvert *et al.* 2004]. Les illusions audiovisuelles telles que l'effet McGurk [1] ou la ventriloquie [2] sont souvent utilisées comme "marqueurs" de l'intégration. Pour la première, un /ga/ visuel (prononcé par un locuteur) est doublé par un /ba/ acoustique. Le mélange de

1. `https://www.youtube.com/watch?v=aFPtc8BVdJk`
2. `https://www.youtube.com/watch?v=EADGMYpUa6I`

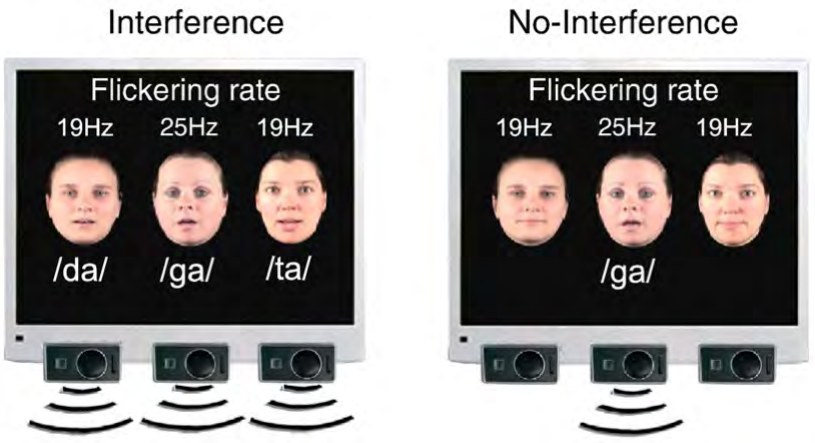


Figure 4.1 – Montage expérimental de l'expérience de Senkowski *et al.*. Trois locuteurs sont présentés sur un écran. Dans la condition "interference" les trois prononcent une syllabe simultanément. Dans la condition "no-interference", seul le locuteur central parle. La tâche est de détecter la syllabe /ba/ lorsqu'elle est prononcée par le locuteur central. Ce dernier est présenté à 25 Hz, les deux autres à 19 Hz. Extrait de [Senkowski *et al.* 2008].

ce stimulus bimodal incongruent est alors perçu par les participants comme étant un /da/ [McGurk & MacDonald 1976]. Pour la seconde, la localisation (temporelle ou spatiale) apparente d'un événement dans une modalité donnée est déplacée vers un événement concurrent appartenant à autre modalité [Thurlow & Jack 1973]. Dans un cas comme dans l'autre, plus l'intégration fonctionne, plus le biais est fort.

La robustesse de l'intégration audiovisuelle est modulée entre autres par la nature des stimuli présentés. Elle est particulièrement efficace lorsqu'il s'agit de locuteurs et de sons de parole (voir discussion section 3.4). Cette robustesse est telle que l'intégration semble opérer de manière automatique, pourvu que les caractéristiques bas niveau des stimuli bimodaux soient suffisamment corrélées [Green *et al.* 1991, Munhall *et al.* 1996]. Mais si tel était le cas, comment ferions-nous pour discriminer les "bonnes" co-occurences intermodales (celles bien liées au même événement) des simples coïncidences (issues de sources indépendantes) ? De nombreuses expériences ont mis en évidence le rôle décisif de l'attention dans ce processus ([Driver & Spence 1998,Bertelson & de Gelder 2004,Tiippana *et al.* 2004], voir [Navarra *et al.* 2010] pour un état de l'art). Par exemple, il a été montré que si l'on distrait visuellement des sujets durant une expérience type McGurk, la prévalence de l'effet est de 30% moindre par rapport à des sujets non distraits [Tiippana *et al.* 2004]. L'importance des distracteurs - et donc de l'attention - dans la perception audiovisuelle de la parole a également été très joliment mise en lumière par Senkowski *et al.* Dans leurs études, ces auteurs ont présenté à leurs sujets trois locuteurs. La tâche était de détecter la syllabe /ba/ prononcée par le locuteur central (cible) alors que les locuteurs périphériques (distracteurs) se taisaient (condition sans intérférence) ou prononçaient en même temps une autre syllabe (condition interférence), comme l'illustre la Figure 4.1 [Senkowski *et al.* 2008]. Dans

le même temps, les auteurs ont manipulé la présentation des visages en les faisant clignoter à différentes fréquences : le locuteur cible était présenté à 25 Hz alors que les distracteurs étaient à 19 Hz. La méthode des *Steady-State Visual Evoked Potentials (SSVEP)* a été utilisée pour mesurer en temps réel vers quel locuteur se portait l'attention visuelle (*overt* ou *covert*) des sujets à partir de leurs signaux EEG. Les résultats montrent que l'amplitude des potentiels évoqués par les distracteurs est corrélée négativement avec la performance des sujets : le déploiement de l'attention visuelle vers les locuteurs distracteurs interfère avec la perception audiovisuelle du discours du locuteur cible. Dans cette étude, l'attention visuelle module l'intégration. Mais qu'en est-il de l'attention auditive ? Pour tenter de répondre à cette question, Alsius & Soto-Faraco ont mené deux expériences [Alsius & Soto-Faraco 2011]. Lors de la première, les sujets devaient détecter quelle face, parmi un ensemble de visages distracteurs, était à l'origine du discours prononcé (Figure 4.2a). Dans la seconde, au contraire, les participants devaient détecter quel signal de parole, parmi un ensemble de discours distracteurs, coïncidait avec le visage qui leur était présenté (Figure 4.2b). Les résultats montrent que dans la première expérience, les temps de réaction augmentent avec le nombre de visages à discriminer : l'"appariement" audiovisuelle est ici une tâche sérielle nécessitant le déploiement de l'attention spatiale visuelle. A l'inverse, dans la seconde expérience, les temps de réaction sont indépendants du nombre de flux sonore à discriminer : l'extraction de correspondances audiovisuelles dans les signaux de parole est effectuée en parallèle, sans déploiement de l'attention spatiale. Les auteurs soulignent que ces résultats révèlent une différence fondamentale entre perception visuelle et auditive. Alors qu'il est nécessaire d'encoder la position spatiale d'un objet visuel pour avoir accès à ses caractéristiques (couleurs, forme...), ce n'est pas le cas pour un objet sonore : on peut percevoir l'intensité ou la fréquence d'un son en ignorant tout de la localisation de sa source.

Ces études indiquent que l'attention joue un rôle fondamental pour l'intégration audiovisuelle de la parole. En cela, elles sont en contradiction avec d'autres recherches menées à partir de stimuli artificiels. Par exemple, le *pip and pop phenomenon*, dont nous avons déjà parlé, a mis en évidence une intégration automatique, en parallèle d'une cible visuelle avec un *pip* sonore (voir section 1.4.2.2 [Van der Burg *et al.* 2008]). Afin de rendre compatibles ces résultats a priori contradictoires, Talsma *et al.* ont proposé un modèle dans lequel la complexité du stimulus joue un rôle clef dans la nature de l'intégration audiovisuelle [Talsma *et al.* 2010]. Dans leur modèle, l'intégration multimodale s'effectue de manière pré-attentive, en parallèle, dans une scène où la compétition entre les stimuli est faible. Dans le *pip and pop phenomenon*, le très saillant *pip* sonore générerait un signal suffisamment fort pour être automatiquement associé au stimulus concomitant dans la modalité visuelle. A l'inverse, lorsque dans chaque modalité de multiples stimuli sont en compétition et qu'aucun ne ressort particulièrement de la scène (comme c'est le cas pour la perception audiovisuelle de la parole dans les expériences d'Alsius & Soto-Faraco), l'attention *top-down* des observateurs est requise pour associer les sti-

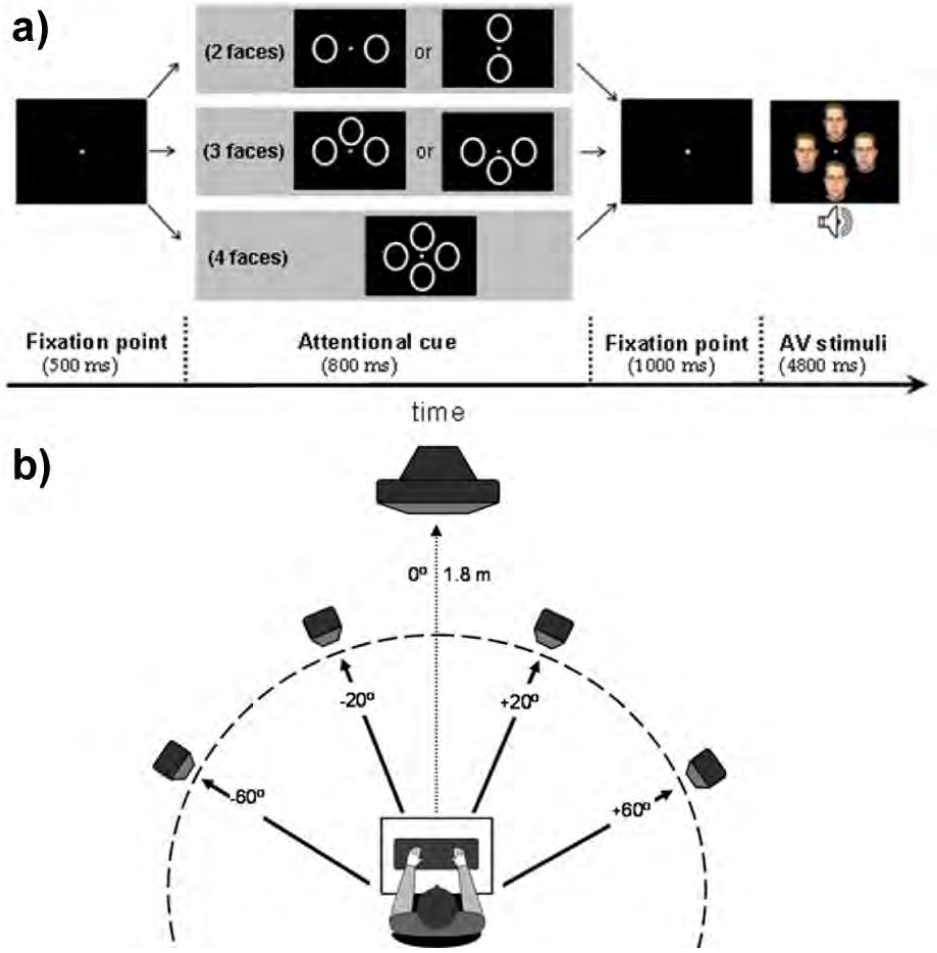


Figure 4.2 – Montages expérimentaux des expériences d'Alsius & Soto-Faraco. **(a)** Point de fixation central, suivi d'un indice spatial, puis du stimulus constitué d'un point de fixation central et de quatre visages, chacun prononçant un discours différent, ainsi que d'un unique signal de parole. Les participants devaient détecter la tête correspondant au son diffusé. **(b)** Les signaux de parole étaient diffusés par des haut-parleurs placés aux positions indiquées. Les sujets regardaient le centre de l'écran où était présentée une tête parlante, et devaient détecter le haut-parleur lui correspondant. Adapté de [Alsius & Soto-Faraco 2011].

muli pertinents (voir aussi la théorie de la charge perceptuelle de Lavie [Lavie 2005]).

Si la corrélation entre les attributs bas niveau est depuis toujours invoquée comme condition nécessaire à l'intégration audiovisuelle, l'attention semble aussi être une variable décisive lors de la perception audiovisuelle de la parole. Et pour étudier les processus attentionnels, quoi de mieux que les mouvements oculaires ?

4.1.3 Visages, parole, et mouvements oculaires

Dans un premier temps, conformément aux modèles de saillance visuelle classiques, on pourrait penser que notre regard est attiré par la zone la plus saillante du visage du locuteur : sa bouche en mouvement. Or nous avons vu dans la section précédente que les stratégies d'exploration des visages dépendent de nombreuses variables (contexte, tâche, intention de l'observateur...). Afin d'identifier et de quantifier ces différentes stratégies en fonction de l'intelligibilité du discours, Vatikiotis-Bateson *et al.* ont présenté des visages prononçant un monologue en faisant notamment varier le niveau de bruit acoustique [Vatikiotis-Bateson *et al.* 1998]. Leurs résultats indiquent que les deux principaux attracteurs de regard sont la bouche et les yeux. Le temps de regard dédié à la bouche augmente avec le niveau de bruit, n'excédant toutefois jamais la moitié du temps total d'exploration. Cette propension à continuer à fixer les yeux même à de hauts niveaux de bruit est expliquée d'une part par la large distribution de l'information phonétique (non limitée à la région de la bouche), et d'autre part par la capacité des sujets à percevoir ces gestes oro-faciaux en vision périphérique [Bailly *et al.* 2012]. Ces résultats sont en accord avec ceux de Paré *et al.* qui ont montré que l'effet McGurk persiste lorsqu'on force le regard à s'éloigner de la bouche, et ne devient négligeable que pour une excentricité de regard de 60 ° [Paré *et al.* 2003]. Ces résultats démontrent que l'analyse des hautes fréquences spatiales fournies par la vision fovéale directe n'est pas nécessaire au bon traitement de l'information visuelle du discours. L'adaptation des stratégies d'exploration du visage en fonction de l'information recherchée a également été mise en évidence par des études demandant aux sujets de juger de l'état émotionnel du locuteur ou de reconnaître les mots prononcés à différents niveaux de bruit [Lansing & McConkie 2003, Buchan *et al.* 2007]. Ces études ont montré que lors d'une tâche de jugement émotionnel, les yeux sont davantage fixés que la bouche, et que l'inverse se produit si les sujets tentent de comprendre ce qui est dit. Lorsque le rapport signal sur bruit se dégrade, les sujets semblent se concentrer sur le milieu du visage : ils fixent moins les yeux et accroissent leur durée de fixation sur le nez et la bouche. Ces résultats sont en accord avec ceux de Vatikiotis-Bateson *et al.* qui ont montré que le nombre de transitions entre les différentes parties du visage (bouche, nez, yeux) diminue en présence de bruit. Ces études montrent que la qualité de l'information verbale affecte la manière dont le regard l'acquiert. Mais que se passe-t-il si l'on supprime totalement cette dernière ? C'est ce qu'ont regardé Võ *et al.* en présentant des vidéos de personnes interviewées

(a) **(b)**

Figure 4.3 – **(a)** Stimulus utilisé par Hirvenkari et collègues, adapté de [Hirvenkari *et al.* 2013]. **(b)** Stimulus utilisé par Foulsham & Sanderson, extrait de [Foulsham & Sanderson 2013]. Un participant regarde un locuteur donné dès lors que son regard tombe dans le rectangle correspondant.

avec ou sans leurs bande-son [Võ *et al.* 2012]. La suppression de l'information sonore a réduit la proportion de fixations sur les visages en général (au profit de l'arrière plan) et sur la bouche en particulier (au profit des yeux et du nez). Ensemble, ces résultats indiquent que l'exploration d'un visage en train de parler n'obéit pas à des règles universelles mais dépend bien du type d'information disponible et / ou recherchée.

Les recherches évoquées jusqu'à présent utilisent des stimuli audiovisuels constitués d'un visage en gros plan pour l'image et d'un signal de parole pour le son. Ce dispositif expérimental, s'il a l'avantage d'être facilement contrôlable et comparable n'est toutefois que peu représentatif des situations que nous expérimentons dans la vie de tous les jours. En effet, une conversation est souvent composée de plusieurs locuteurs, généralement plongés dans une scène complexe (objets, arrière-plan), et ne se contentant pas d'écouter ce qu'il se dit mais interagissant, prenant la parole de manière dynamique. L'étude du "regard social" (*social gaze*) est en plein développement. Foulsham *et al.* ont montré que lorsque l'on regarde un groupe essayant de prendre une décision, notre regard est attiré vers la personne ayant la parole, et ce d'autant plus que cette dernière est jugée comme ayant un statut social "dominant" [Foulsham *et al.* 2010]. Lors d'une conversation, nous produisons ou interprétons sans cesse des "indices de regard" afin de clarifier un discours ambigu [MacDonald & Tatler 2013]. En effet, sortie de son contexte, la phrase "Passe-moi le truc posé sur le machin, s'il te plaît" est pour le moins équivoque, alors qu'un regard appuyé sur le tire-bouchon la rend parfaitement compréhensible. Il a par ailleurs été montré que lorsque deux personnes regardent ou discutent de quelque chose, leurs mouvements oculaires sont très sensibles à ce qu'ils pensent que l'autre voit ou croit [Richardson & Dale 2005, Richardson *et al.* 2012]. Deux récentes études se sont intéressées aux mouvements oculaires effectués lorsqu'on assiste à une conversation entre plusieurs personnes [Hirvenkari *et al.* 2013, Foulsham & Sanderson 2013]. Les auteurs ont présenté à leurs participants des vidéos de deux ou quatre personnes tenant une conversation (Figure 4.3) dans trois conditions : avec le son et l'image (AV), sans le son (V), ou avec une frame fixe (A). Logiquement,

Table 4.1 – Proportions des fixations atterrissant sur le locuteur, un auditeur, ou sur le reste de l'image. Dans l'étude de Hirvenkari *et al.*, les conversations étaient tenues par 2 personnes (Figure 4.3a), et présentées dans les conditions audiovisuelle (AV, son + image), purement visuelle (V, sans son) ou purement auditive (A, image fixe). Dans l'étude de Foulsham & Sanderson, les conversations étaient tenues par 4 personnes (Figure 4.3b), et présentées dans les conditions audiovisuelle ou purement visuelle.

	Hirvenkari *et al.*			**Foulsham *et al.***	
Conditions	**AV**	**V**	**A**	**AV**	**V**
Fixations sur le locuteur (%)	72	69	63	51	37
Fixations par auditeur (%)	23	27	25	15	20
Fixations sur le reste (%)	5	4	12	4	3

les deux études ont montré que quelle que soit la condition, les visages accaparent presque tous les regards (Table 4.1). De plus, les prises de parole ont une grande influence sur la façon dont la conversation est regardée : un locuteur attire bien plus le regard qu'un auditeur, particulièrement dans la condition audiovisuelle. Cependant, dans les conditions unimodales, les observateurs suivent aussi (même si moins précisément) les tours de parole, ce qui indiquent que les informations visuelles et auditives sont au moins en partie redondantes. Foulsham & Sanderson ont également montré que la condition sonore ne modifiait pas les régions du visage fixées par les participants : avec ou sans son, les yeux (38%) restent davantage regardés que la bouche (15%).

Aucune des études mentionnées jusqu'à présent ne quantifient précisément la propension des différents attributs visuels à attirer le regard dans une scène de conversation. Si Birmingham *et al.* ont montré que les modèles classiques de saillance n'expliquaient en rien les fixations dans des scènes sociales statiques, qu'en est-il pour les scènes dynamiques, où le mouvement est connu pour être un puissant attracteur d'attention ? Pour répondre à cette question, dans un premier temps nous allons développer l'analyse des résultats de la catégorie Visages obtenus lors de l'expérience 2 et décrits au Chapitre 3. Puis, nous appliquerons les techniques statistiques présentées section 3.3 en prenant en compte les visages parmi les paramètres de modélisation, et estimerons l'influence de la condition expérimentale et des tours de parole sur les valeurs relatives de leurs coefficients.

4.2 Expérience 2, catégorie Visages

Lors de cette expérience, les vidéos de la catégorie Visages ont été présentées dans quatre conditions sonores. Originale : avec la bande-son de la conversation originale ; Mix Intra : avec la bande-son d'une autre vidéo de la même catégorie ; Son Paysages : avec la bande-son d'une vidéo de la catégorie Paysages (bruit de la pluie, de la mer, du vent dans les feuilles...) et Son POM : avec la bande-son d'une vidéo de la catégorie Plusieurs Objets en Mouvement (bruits brefs d'objets tombant, s'entrechoquant...).

4.2.1 Stimuli et segmentation des visages

Les 15 vidéos de la catégorie Visages sont des plans uniques extraits de films commerciaux francophones. Chaque scène présente 2 à 4 personnes ayant une conversation dans un environnement naturel (café, rue, couloir, bureau...), entourées de divers objets, éventuellement en mouvement (papiers, cigarettes, voitures...). Les vidéos durent entre 12 et 30 secondes (M = 19.6 s ; SD = 4.9 s). Pour marquer les zones de l'image correspondant aux visages des différents locuteurs, nous avons utilisé Sensarea, un programme conçu au laboratoire par Pascal Bertolino permettant de segmenter automatiquement ou semi-automatiquement des objets dans des vidéos [Bertolino 2012]. Ici, les visages sont repérés sur chaque frame par un masque ovale (Figure 4.6a). L'aire moyenne des visages est de $3.3 \pm 0.4 \times 5.2 \pm 0.9$ deg^2, et l'espacement moyen est de 10 ± 2 deg. Ainsi, en moyenne, chaque visage n'occupe que (3.3 × 5.2) / (28 × 22.5) = 2.7 % de l'aire totale de l'image. Afin d'étudier l'influence des tours de parole, nous avons manuellement repéré pour chaque visage les périodes temporelles de parole et de silence. Ces périodes temporelles sont définies à partir de l'information sonore contenue dans les bandes-son originales. Pour un total de 33 visages, 28 parlent au moins une fois et 27 sont silencieux au moins une fois.

4.2.2 Résultats

Les deux principaux résultats mis en évidence au chapitre précédent (section 3.2.2.2) sont

1. une dispersion plus faible entre les positions oculaires des différents observateurs dans la condition Originale que dans les autres conditions, et
2. des amplitudes de saccade médianes plus courtes dans la condition Originale que dans les autres conditions.

Pour mieux interpréter ces derniers, regardons de plus près les distributions d'amplitude de saccade dans les différentes conditions expérimentales.

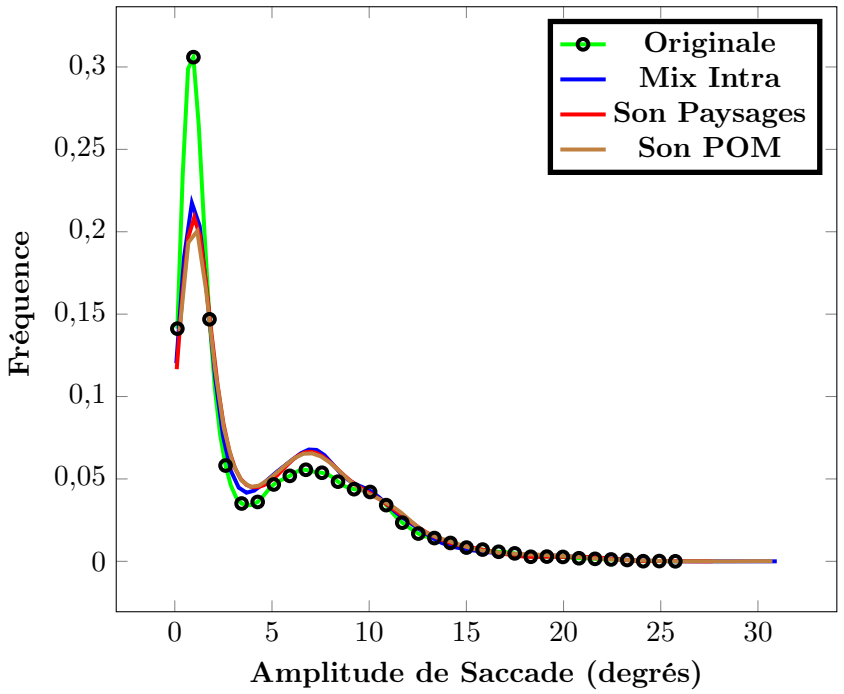


Figure 4.4 – Distributions des amplitudes de saccade médianes de la catégorie visuelle Visages selon la condition expérimentale. Seule la distribution dans la condition Originale (en vert avec un marqueur circulaire) se distingue des autres. Les densités de probabilité ont été estimées sur 100 points régulièrement espacés couvrant l'ensemble des données (fonction Matlab ksdensity).

4.2.2.1 Amplitudes des saccades

La Figure 4.4 représente les distributions d'amplitude de saccade selon la condition expérimentale. Leur forme globale est bien différente que celle dans les autres catégories visuelles (Figure 3.3). Nous sommes en présence de distributions bimodales, avec un mode principal vers 1° et un mode secondaire vers 7°. Si les distributions des trois conditions Non Originales (Mix Intra, Son Paysages et Son POM) se confondent, celle de la condition Originale se distingue par un mode principal plus prononcé et un mode secondaire légèrement plus aplati (3 tests de Kolmogorov-Smirnov entre la condition Originale et les 3 autres conditions, tous les $p < .001$). Pour mieux comprendre l'origine de ces deux modes, nous avons séparé les saccades en deux groupes : les saccades courtes (moins de 3°) correspondant au mode principal, et les saccades longues (plus de 3°) correspondant au mode secondaire. Dans chaque groupe, nous avons comparé la proportion de saccades (1) partant d'un visage et arrivant sur un autre (Inter Visages), (2) partant d'un visage et arrivant sur le même (Intra Visage) et (3) partant ou arrivant d'ailleurs qu'un visage (Autre). Comme le montre la Figure 4.5, il n'y a pas de saccades Inter Visages appartenant au mode principal, et quasiment pas de saccades Intra Visages appartenant au mode secondaire. Nous en concluons que le mode principal représente les petites saccades

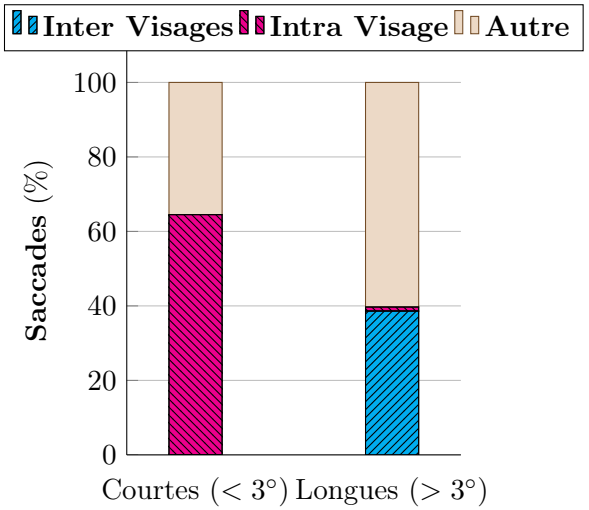


Figure 4.5 – Proportion de saccades toutes conditions expérimentales confondues (1) partant d'un visage et arrivant sur un autre (Inter Visages), (2) partant d'un visage et arrivant sur le même (Intra Visage) et (3) partant ou arrivant d'ailleurs qu'un visage (Autre). Les saccades sont séparées en deux groupes : les saccades courtes ($< 3°$), correspondant au mode principal des distributions de la Figure 4.4, et les saccades longues ($> 3°$), correspondant au mode secondaire.

faites à l'intérieur d'un même visage (nez, bouche, yeux...) alors que le mode secondaire représente les saccades effectuées pour passer d'un visage à un autre.

4.2.2.2 Taux de fixations par visage

Dans la condition expérimentale Originale, les participants font donc davantage de petites saccades à l'intérieur d'un même visage que dans les conditions Non Originales. Pour comprendre ce phénomène, nous avons calculé frame par frame les taux de fixations atterrissant dans chaque visage (nombre de fixations sur le visage divisé par le nombre total de fixations sur la frame). Pour prendre en compte l'influence des tours de parole, nous avons pris la moyenne de ces taux sur les périodes correspondant à l'élocution ou au silence des visages correspondant (Table 4.2). Quelque soit la condition expérimentale, les visages parlants attirent environ deux fois plus de fixations que les visages silencieux. Nous avons mené une ANOVA à un facteur intra (les taux de fixations sur les locuteurs) et quatre niveaux (les conditions expérimentales). Chaque niveau possède 28 items (les 28 visages parlant au moins une fois). Il existe un effet principal de la condition expérimentale ($F(3,81) = 8.9$, $p < .001$) et des comparaisons post-hoc de Bonferroni montrent que les visages parlant sont davantage regardés dans la condition Originale que dans les trois autres conditions (tous les $p < .001$). Il n'y a pas de différences entre les conditions Non Originales (tous les $p = 1$).
La même ANOVA a été menée sur les taux de fixations sur les 27 visages silencieux,

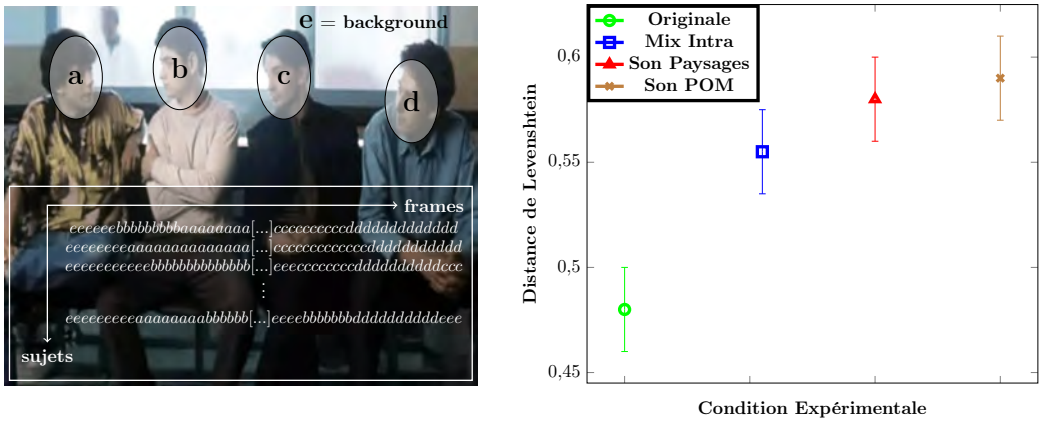
Table 4.2 – Proportions des fixations atterrissant sur le locuteur, un auditeur, ou sur le reste de l'image. Ces ratios sont calculés pour chaque visage de chaque vidéo. Ils sont ensuite moyennés sur les périodes où les visages correspondant parlent ou se taisent ($M(\pm SE)$). La somme d'une colonne n'est pas nécessairement égale à 100% car le nombre de locuteur varie selon les vidéos (voir Annexe B).

	Condition Expérimentale			
	Originale	Mix Intra	Son POM	Son Paysages
Fixations sur le locuteur (%)	48 (±5)	40 (±4)	38 (±4)	38 (±5)
Fixations par auditeur (%)	20 (±4)	23 (±3)	22 (±3)	22 (±3)
Fixations sur le reste (%)	30 (±3)	34 (±3)	37 (±3)	35 (±3)

mais nous n'avons pas trouvé d'effet de la condition expérimentale ($F(3,78) = 1.5$, $p = .21$).
Nous avons également mené cette analyse avec les taux de fixations effectuées sur le reste de l'image (soit 15 items, un par vidéo). Il existe un effet principal de la condition expérimentale ($F(3,42) = 5.9$, $p = .002$), le reste de l'image est moins fixé dans la condition Originale que dans la condition Son Paysages ($p = .03$) et Son POM ($p = .001$). Par contre, il n'y a pas de différence significative avec la condition Mix Intra ($p = .32$). Il n'y a pas non plus de différences entre les conditions Non Originales (Son Paysages *vs.* Mix Intra et *vs.* Son POM, $p = 1$, Mix Intra *vs.* Son POM, $p = .27$).
Les taux de fixations effectuées ailleurs que sur les visages peuvent sembler élevés comparés aux résultats de la littérature (Table 4.1). Il est important de noter qu'à l'inverse des stimuli précédemment utilisés (Figure 4.3), ceux utilisés dans ce travail présentent des conversations se tenant dans des environnements naturels complexes, avec de nombreux objets autres que les visages susceptibles d'attirer le regard des observateurs. Nous savons à présent que les observateurs d'une conversation regardent davantage le locuteur, et ce d'autant plus s'ils ont accès à l'information sonore associée. Pour affiner cette comparaison des stratégies d'exploration visuelle selon la condition expérimentale, nous allons utiliser une mesure quantifiant directement la similarité entre les scanpaths.

4.2.2.3 Distance de Levenshtein

La distance de Levenshtein est une distance initialement conçue pour mesurer la similarité entre deux chaînes de caractères [Levenshtein 1966]. Cette métrique donne le nombre d'opérations minimum nécessaires pour transformer une séquence



(a) (b)

Figure 4.6 – **(a)** l'image est partitionnée en 5 régions d'intérêt (visage a, visage b, visage c, visage d and background e). Une chaîne de caractères représente le scanpath d'un participant : chaque lettre correspond à la région que la personne a regardé durant la frame correspondante. **(b)** : Distance de Levenshtein normalisée entre les scanpaths et la séquence vérité terrain, dans chaque condition expérimentale. Les barres d'erreurs correspondent aux erreurs standards.

en une autre (insertion, suppression ou substitution d'un seul caractère). Elle a été fréquemment utilisée pour comparer la similarité entre différents scanpaths. Dans ce cas, l'image observée est séparée en différentes régions, et à chacune est attribué un caractère. Les observateurs les explorent de manière séquentielle, et l'ordre de fixation défini une chaîne de caractères propre à chaque participant. Pour une description approfondie des différentes métriques utilisées pour comparer les scanpaths, voir [Le Meur & Baccino 2013]. Ici nous avons utilisé une métrique de base car notre but est simplement de comparer la dynamique de fixation de quelques régions d'intérêt (les visages et le fond), sans tenir compte de la distance les séparant. Pour chaque vidéo, nous avons échantillonné le scanpath de chaque sujet avec une fréquence d'une frame. Nous avons attribué à chaque position oculaire une lettre correspondant à la région fixée (visage a, visage b, visage c, visage d and background e, voir Figure 4.6a). Nous avons également défini une "séquence Vérité Terrain" (**VT**) pour chaque vidéo. Si une vidéo comporte m frames, alors **VT** est un vecteur de taille m tel que si le visage a parle à la frame i, alors $\mathbf{VT}(i) = a$. Si aucun visage ne parle à la frame j alors $\mathbf{VT}(j) = background$. Cette définition est conservatrice dans la mesure où d'habitude, lorsque dans une conversation personne ne parle, les observateurs regardent tout de même les visages. Pour chaque sujet, nous avons calculé la distance de Levenshtein normalisée moyenne entre ses scanpaths sur chaque vidéo et les **VT** correspondantes. La distance de Levenshtein normalisée est la distance de Levenshtein divisée par le nombre de frames m de la vidéo. Nous avons mené une ANOVA à un facteur intra (la condition expérimentale) sur la distance de Levenshtein normalisée moyenne. Chaque niveau a 72 items (les participants). Il existe un effet principal de la condition expérimentale ($F(3,213) = 17.6$, $p < .001$), la distance de Levenshtein normalisée entre **VT** et

les scanpaths enregistrés dans la condition Originale est plus petite que celles enregistrées dans les 3 conditions Non Originales (tous les $p < .001$). Nous n'avons pas trouvé de différences entre les conditions Non Originales (Son POM *vs.* Mix Intra, $p = .12$, Son Paysages *vs.* Son POM, $p = 1$, Mix Intra *vs.* Son Paysages, $p = .64$).

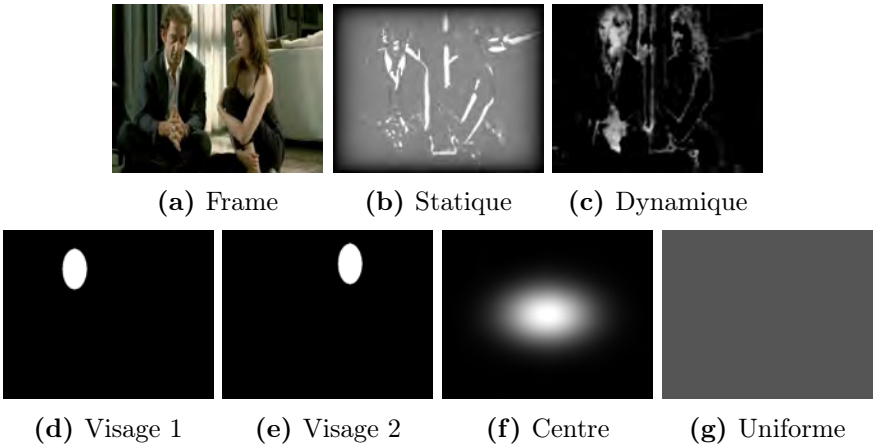
Nous avons montré que dans une scène naturelle dynamique, les visages attirent la grande majorité des fixations. Les locuteurs attirent particulièrement les regards : quelle que soit la condition expérimentale, ils ont environ deux fois plus de chance d'être fixés qu'une personne silencieuse. Dans la condition Originale, les participants passent moins d'un visage à l'autre, mais effectuent plus de courtes saccades au sein d'un même visage. L'analyse temporelle des différentes explorations enregistrées révèle que les observateurs regardent davantage les locuteurs dans la condition Originale. Par contre, nous n'avons trouvé aucune différence entre les conditions Non Originales. Ces résultats sont confirmés par une comparaison fine entre les séquences de fixations et la dynamique de la conversation. Nous avons trouvé une plus grande similarité de la séquence vérité-terrain avec les scanpaths enregistrés dans la condition Originale qu'avec ceux enregistrés dans les autres conditions.
Mais pourquoi regarde-t-on tant le visage les locuteurs ? Sommes nous attirés par des facteurs haut niveau (le locuteur est la principale source d'information sémantique) ? Par des facteurs bas niveau (le locuteur bouge, et le mouvement attire l'attention) ? Ou d'une combinaison de ces derniers ? Pour répondre à ces questions, nous allons utiliser les outils statistiques présentés au chapitre précédent.

4.3 Modélisation statistique

Nous souhaitons quantifier l'importance relative de différents attributs visuels pour expliquer les cartes de densité de positions oculaires que nous avons enregistrées selon la condition expérimentale. Dans ce but, nous utilisons l'algorithme Lasso dans le cadre théorique dressé section 3.3.

4.3.1 Estimation du poids des attributs

Nous reprenons les attributs visuels décrits section 3.3.2 (saillance statique, saillance dynamique, biais de centralité, carte uniforme), auxquels nous ajoutons des "cartes visage" représentant les visages présents à l'image (Figure 4.7). Les cartes visage sont construites à partir des masques binaires créés grâce au logiciel Sensarea (Figure 4.7d et 4.7e). Les poids de chaque attribut sont estimés pour chaque frame de chaque vidéo. Les poids par vidéo sont calculés en moyennant sur l'ensemble des frames. Figure 4.8a sont représentés les poids moyens sur les 15 vidéos de la catégorie Visages, selon la condition expérimentale.



(a) Frame **(b)** Statique **(c)** Dynamique

(d) Visage 1 **(e)** Visage 2 **(f)** Centre **(g)** Uniforme

Figure 4.7 – **(a)** Frame extraite d'une vidéo de la catégorie Visages. Attributs utilisés pour modéliser la carte de densité des positions oculaires associée : **(b)** saillance statique, **(c)** saillance dynamique, **(d)** et **(e)** une carte pour chaque visage, **(f)** biais de centralité et **(g)** carte uniforme.

Nous avons mené une ANOVA à deux facteurs intra (les attributs et la condition expérimentale). Il existe un effet principal des attributs ($F(4,56) = 145.95$, $p < .001$), le poids moyen des Visages étant supérieur à tous les autres (tous les $p < .001$). Nous n'avons trouvé aucune différence entre les poids de la saillance statique, de la saillance dynamique et du biais de centralité (Statique *vs.* Dynamique, $p = 1$, Statique *vs.* Centre, $p = .67$, Dynamique *vs.* Centre, $p = 1$). Le poids moyen de la carte uniforme est inférieur aux poids de la saillance statique et de la saillance dynamique (Uniforme *vs.* Statique, $p < .001$, Uniforme *vs.* Dynamique, $p = .06$), mais n'est pas significativement différent du biais de centralité (Uniforme *vs.* Centre, : $p = .18$). Il existe également un effet principal de la condition expérimentale ($F(3,42) = 74.39$, $p < .001$) et de l'interaction ($F(12,168) = 6.43$, $p < .001$). Des comparaisons post-hoc de Bonferroni entre les conditions expérimentales ont été calculées pour chaque attribut. Nous n'avons trouvé aucune différence entre les conditions expérimentales pour les attributs saillance statique, saillance dynamique, biais de centralité et carte uniforme (tous les $p = 1$). Par contre, le poids moyen de l'attribut Visages est plus élevé dans la condition Originale que dans les autres conditions (Originale *vs.* Mix Intra $p = .019$; Originale *vs.* Son POM $p < .001$; Originale *vs.* Son Paysages $p < .001$). Nous n'avons trouvé aucune différence entre les poids Visages dans les conditions Non Originales (Son POM *vs.* Mix Intra $p = .32$; Son Paysages *vs.* Son POM $p = 1$; Mix Intra *vs.* Son Paysages $p = 1$).

4.3.2 Visages parlants et visages silencieux

Pour séparer les contributions des visages parlants des visages silencieux, nous avons moyenné les poids de chacun des visages sur les mêmes périodes temporelles

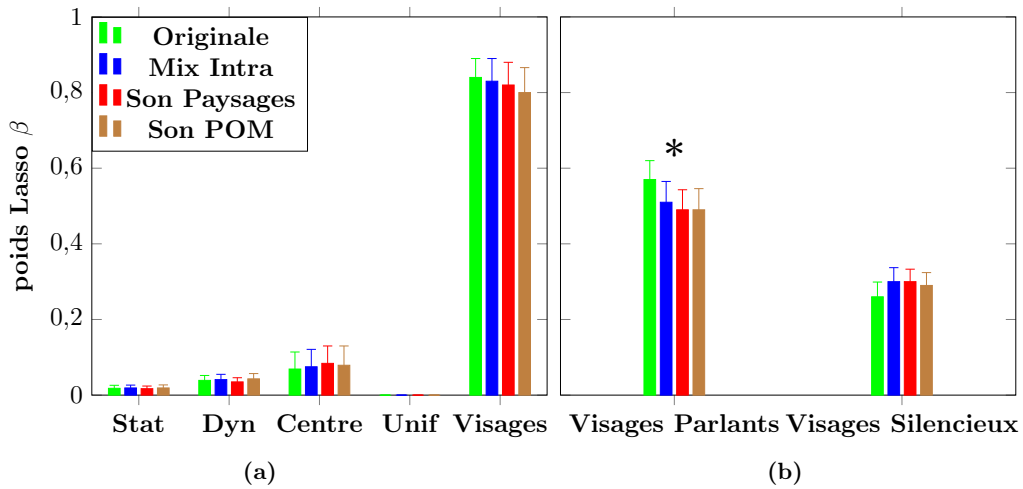


Figure 4.8 – **(a)** Valeurs moyennes des poids des attributs choisis pour modéliser les cartes de densité de positions oculaires ($\sum_{i=1}^{5} \beta_i = 1$). **(b)** Contributions des visages parlants et des visages silencieux à l'attribut "Visages" ($\beta_{\text{Parlants}} + \beta_{\text{Silencieux}} = \beta_{\text{Visages}}$). Les poids sont moyennés sur toutes les frames de chaque vidéo, puis sur l'ensemble des vidéos. Les barres d'erreurs correspondent aux erreurs standards.

que celles définies pour le calcul des taux de fixations par locuteur et par auditeur (section 4.2.1). Ces poids (Figure 4.8b) sont du même ordre que les taux de fixations reportés Table 4.2 : autour de 20% pour les visages silencieux, quelle que soit la condition expérimentale, environ 50% pour les visages parlants dans la condition Originale, et 40% dans les autres conditions.
Nous avons mené une ANOVA à deux facteurs intra : le type de visage (Parlant ou Silencieux) et la condition expérimentale. Il existe un effet principal du type de visage ($F(1,14) = 106.75$, $p < .001$), de la condition expérimentale ($F(3,42) = 5.16$, $p = .004$) et de l'interaction ($F(3,42) = 20.14$, $p < .001$). Le poids moyen des visages parlants est supérieur dans la condition Originale que dans les autres conditions (tous les $p < .001$). La condition expérimentale n'a pas d'effet sur les poids des visages silencieux (Originale *vs.* Mix Intra $p = .70$; Originale *vs.* Son POM $p = 1$; Originale *vs.* Son Paysages $p = 1$; Mix Intra *vs.* Son Paysages $p = 1$; Son POM *vs.* Mix Intra $p = 1$; Son Paysages *vs.* Son POM $p = 1$).

Dans les scènes de conversations, les attributs de saillance bas niveau (tant statique que dynamique) et le biais de centralité n'expliquent pas de manière satisfaisante les positions oculaires d'éventuels observateurs. A l'inverse les visages, et plus spécialement les visages des locuteurs, accaparent l'essentiel de l'attention. Même si la présence du signal de parole original augmente le poids de ces derniers, le regard est principalement dirigé par l'information visuelle. En effet, même avec des bandes-son non originales (incongruentes), le poids du visage d'un locuteur vaut encore le double du poids du visage d'un auditeur. Nous n'avons trouvé aucune différence entre les bandes-son non originales.

4.4 Discussion

4.4.1 Les visages accaparent l'attention

La spécificité des visages dans l'exploration visuelle est connue de longue date [Buswell 1935]. De nombreuses études ont montré, d'abord à partir d'images statiques, puis de stimuli dynamiques plus complexes et proches de la réalité, que dans une scène naturelle les visages attirent l'essentiel des fixations, et ce depuis le plus jeune âge [Frank *et al.* 2009, Foulsham *et al.* 2010, Mital *et al.* 2010, Võ *et al.* 2012]. Avec plus de 65% de fixations sur les visages, nos résultats vont largement dans ce sens, alors même que nos stimuli présentent des personnes au sein d'un environnement complexe et dynamique dans lequel de nombreux objets peuvent potentiellement attirer l'attention (café, corridor, rue, bureau...). Ce regroupement des positions oculaires sur les visages entraîne une diminution de la dispersion et une augmentation de la distance au centre comparé à des scènes ne présentant pas de régions d'intérêt aussi fortes, comme les paysages (voir section 3.2.2.1). Cette particularité de l'exploration des visages se retrouve également dans la distribution des amplitudes de saccade. Traditionnellement asymétrique positive [Bahill *et al.* 1975, Tatler *et al.* 2006], cette distribution devient ici bimodale, avec un mode principal autour de 1°d'angle visuel et un mode secondaire autour de 7°. Lorsque nous explorons une scène contenant plusieurs personnes, nous effectuons principalement deux types de saccades : à l'intérieur d'un même visage (intra : yeux - nez - bouche) et entre deux visages différents (inter). Nous avons comparé la proportion de saccades intra et inter visages selon leur amplitude. Presque l'intégralité des saccades intra-visage appartiennent au mode principal, alors que toutes les saccades inter-visages appartiennent au mode secondaire. Ces résultats sont cohérents avec la disposition des visages dans nos stimuli : ils ont une surface intra moyenne de $3° \times 5°$, correspondant au premier mode, et une distance inter moyenne de 10°, correspondant au second mode.

Cette mainmise des visages sur les regards est également reflétée par notre modélisation statistique, dans laquelle le poids des attributs visuels bas niveau (saillance statique, saillance dynamique, biais de centralité) vaut moins du quart du poids attribué aux cartes visages. De précédentes études, travaillant avec des images statiques, ont également mis en avant l'inefficacité des modèles de saillance bas niveau classiques pour prédire les positions oculaires sur certaines scènes, ces derniers ne prenant pas en compte la dimension sociale de l'exploration visuelle [Birmingham & Kingstone 2009, Tatler *et al.* 2011]. Nos résultats confortent ce constat et le généralisent aux scènes dynamiques.

Taux de fixations et modélisation statistique montrent également que les locuteurs sont davantage regardés que les auditeurs. Les indices visuels guidant ce comportement peuvent être de deux ordres : bas et haut niveau. Rappelons nous de la Figure 3.6 du chapitre précédent, montrant un poids nettement plus fort pour la saillance dynamique dans la catégorie Visages que dans les autres catégories. En

faisant l'hypothèse que les locuteurs bougent davantage que les auditeurs (ce qui est raisonnable, comme nous le verrons au chapitre suivant), le mouvement peut être un attribut crédible pour expliquer leur forte saillance. Ensuite, des indices de plus haut niveau peuvent être invoqués, comme les expressions ou le langage corporel adoptés par l'ensemble des personnes prenant part à la conversation [Richardson *et al.* 2008]. Si cette préférence aux locuteurs est avérée quelle que soit la condition expérimentale, elle est d'autant plus vraie dans la condition Originale.

4.4.2 Influence de la bande-son originale

La catégorie Visages était la seule catégorie visuelle de l'expérience 2 à présenter une nette différence entre conditions expérimentales. Dans cette catégorie, la dispersion et les amplitudes de saccade étaient plus grandes avec les bandes-son incongruentes qu'avec la bande-son originale. L'approfondissement de ces résultats nous fournit un faisceau d'indices montrant que l'accès aux sons de parole prononcés par les personnes présentes à l'écran permet (ou donne envie) aux observateurs de suivre de plus près la distribution des tours de parole. Les plus petites dispersions et amplitudes de saccade mesurées dans la condition Originale ont une même source : les positions oculaires des observateurs sont rythmées par les prises de parole. Ces derniers effectuent plus de petites saccades sur les locuteurs, afin de s'aider des nombreux gestes faciaux non verbaux permettant de mieux saisir leurs propos et émotions [Vatikiotis-Bateson *et al.* 1998, Buchan *et al.* 2007, Võ *et al.* 2012]. A l'inverse, les observateurs ne disposant pas de l'information sonore originale ont du mal à suivre la scène et effectuent davantage de grandes saccades d'un visage à l'autre. Ceci se traduit aussi dans notre modélisation statistique par un plus fort poids estimé pour les locuteurs dans la condition Originale que dans les autres conditions. Ces résultats sont en accord avec ceux publiés dans deux récentes études [Hirvenkari *et al.* 2013,Foulsham & Sanderson 2013], qui ont également noté une corrélation temporelle entre les sons de parole et le déploiement de l'attention sur le visage les prononçant. Ces deux travaux indiquent qu'avec la bande-son originale, le taux de fixations sur le locuteur augmente brutalement au début de la prise de parole, et atteint son maximum après 800 ms à 1 s. Sans bande-son, l'allure générale des scanpaths demeure inchangée, mais la vitesse à laquelle les observateurs vont fixer un visage venant de prendre la parole diminue. Ceci laisse penser que les sons de parole peuvent agir comme un signal indiquant qu'un nouveau locuteur s'exprime : la modalité sonore rythme précisément la dynamique de l'exploration visuelle.

4.4.3 Influence des autres bandes-son

Etonnamment, quelque soit la métrique utilisée, nous n'avons trouvé aucune différence entre les conditions expérimentales non originales (Mix Intra, Son Paysages, Son POM). Nous interprétons cette absence de différence en terme de "liage" audio-

visuel. La fusion audiovisuelle peut être conçue comme un processus en deux étapes : si les signaux unimodaux (par exemple le mouvement de lèvres et le son qu'elles prononcent) présentent une corrélation spatiotemporelle (bas niveau) suffisante, ils sont liés par les observateurs. Puis, sous réserve de ce précoce liage, ces signaux pourront être intégrés ensemble en un percept unifié [Berthommier 2004]. Une récente étude menée au laboratoire est venue renforcer cette conception, en montrant qu'il était possible de "délier" un flux sonore d'un flux visuel, prévenant ainsi leur intégration [Nahorna *et al.* 2012]. Les auteurs se sont servi de l'effet McGurk comme d'un marqueur de l'intégration audiovisuelle : plus cet effet est présent, plus l'intégration est efficace. Ils ont montré que si un stimulus de type McGurk (une personne prononçant un /ga/ visuel doublé par un /ba/ sonore) est précédé d'un contexte incohérent, les probabilités d'occurrence de l'effet McGurk (perception de /da/), et donc l'intégration audiovisuelle, sont largement réduites. Par "contexte incohérent", les auteurs entendent une décorrélation évidente entre les sons prononcés dans le stimulus et les sons entendus par les participants. Ce "débranchement" de l'effet McGurk peut même être causé par un contexte incohérent très court, de l'ordre d'une syllabe.

Psychocinématique 2

Il est intéressant de constater que le concept développé par Nahorna *et al.* a depuis longtemps été compris et exploité par les cinéastes. Ainsi, le compositeur et théoricien du cinéma Michel Chion réfute la notion de bande-son en tant qu'unité cohérente en ces termes [a] :

> *En formulant qu'il n'y a pas de bande-son, nous voulons donc dire, pour commencer, que les sons du film ne forment pas, pris à part de l'image, un complexe en soi doté d'une unité interne, qui se confronterait globalement à ce qu'on appelle la bande-image. Mais aussi, nous voulons dire que chaque élément sonore noue avec les éléments narratifs contenus dans l'image - personnages, action - ainsi qu'avec les éléments visuel de texture et de décor, des* rapports verticaux simultanés *bien plus directs, forts et prégnants que ceux que ce même élément sonore peut nouer parallèlement avec les autres sons, ou que les sons nouent entre eux dans leur succession. C'est comme une recette : auriez-vous mélangé à part les constituants sonores avant de les verser sur l'image, qu'une réaction chimique se produira qui désolidarisera les sons et les fera réagir chacun individuellement au champ visuel.*

Michel Chion, Olha Nahorna et ce travail s'accordent sur la nécessité pour un son d'*entrer en rapport vertical simultané* (en d'autres termes, d'être corrélé) avec l'information visuelle afin qu'ils puissent être liés puis intégrés ou, selon les mots du théoricien du cinéma, afin qu'ils puissent *réagir* ensemble.

a. *L'audio-vision : Son et image au cinéma*, éditions Nathan, 1990

Vus depuis notre étude, ces résultats peuvent signifier que l'absence de différence entre les trois conditions expérimentales incongruentes est due à une trop forte décorrélation entre les bandes-son et le contenu visuel associé. Faute de passer la première étape de liage, les bandes-son ne seraient pas intégrées au flux visuel. En d'autres termes, les participants pourraient simplement faire abstraction d'une information auditive déliée n'apportant aucune information complémentaire à l'information visuelle, et se concentrer sur cette dernière. Ainsi, une bande-son incongruente produirait le même effet que pas de bande-son du tout, à savoir une exploration uniquement guidée par la modalité visuelle. Cette interprétation est cohérente avec les résultats de l'expérience 1 (section 2.1.4.1) et avec ceux présentés dans [Foulsham & Sanderson 2013], où était comparée l'exploration visuelle de vidéos avec ou sans leurs bandes-son originales. En effet, les différences constatées entre les conditions "avec" et "sans son" ou "avec bandes-son originales" et "avec bandes-son incongruentes" vont dans le même sens. Nous avons par exemple mesuré dans les deux cas une baisse de la dispersion entre les positions oculaires avec les bandes-son originales.

Notre étude montre donc qu'au sein d'un environnement complexe et dynamique, les visages, et particulièrement les locuteurs, agissent comme des trous noirs perceptifs, attirant vers eux l'essentiel des regards. Quels sont les indices audiovisuels guidant les observateurs vers un visage plutôt qu'un autre ? Comment modéliser computationnellement ce comportement ? Nous tenterons de répondre à ces questions dans le prochain chapitre.

Chapitre 5

Modèle de saillance audiovisuelle pour des scènes de conversation

Sommaire

5.1 Introduction

Les modèles de saillance visuelle "classiques" ne prennent pas en compte de nombreux aspects de la perception visuelle, comme son caractère social, ou sa forte dépendance au contexte général de la scène. De fait, ils ont de faibles performances dès qu'il s'agit de modéliser l'attention visuelle pour des scènes à fort contenu sémantique [Nyström & Holmqvist 2008, Tatler *et al.* 2011]. Dans [Birmingham & Kingstone 2009], les auteurs montrent qu'alors que les régions "sociales" d'une scène statique (les yeux en l'occurence) sont très rapidement fixées (moins de 200 ms), les modèles de saillance classiques, comme celui d' [Itti *et al.* 1998], ne les mettent pas en évidence.
Afin de pallier ce manque, certains auteurs proposent de renforcer la saillance des régions correspondant aux visages présents à l'image. Un attribut haut niveau "visage" est alors ajouté aux attributs bas niveau classiques, améliorant considérablement la performance de leurs modèles [Chen *et al.* 2003, Ma *et al.* 2005, Cerf *et al.* 2008a, Cerf *et al.* 2008b, Goferman *et al.* 2012, Marat *et al.* 2013]. L'intégration des visages dans

une carte de saillance maîtresse est un problème épineux, ces derniers ayant une distribution spatiale et une plage dynamique souvent très différentes des autres attributs. Différentes approches ont été proposées dans la littérature, allant de la simple moyenne [Cerf *et al.* 2008b] à une pondération dépendant de l'indice de confiance de détection des visages [Marat *et al.* 2013], ou encore de l'excentricité, du nombre, de la taille, et de la position relative des visages [Rahman *et al.* 2014]. Cependant, l'ensemble de ces travaux n'exploitent que des attributs visuels, mettant de côté les informations issues d'autres modalités, comme le signal sonore.

Problématique

Comme nous l'avons mis en évidence dans le chapitre précédent, le son joue un rôle significatif dans l'exploration visuelle de scènes de conversation. Avec la bande-son originale, les observateurs suivent de plus près les tours de parole, s'attardant davantage sur le visage des locuteurs. Dans ce chapitre, nous cherchons à détecter automatiquement ces derniers afin de renforcer leur saillance et d'améliorer la prédiction des positions oculaires, par rapport à un modèle donnant une importance égale et constante à tous les visages présents à l'image. Cependant, nous avons vu que même avec une bande-son incongruente, les observateurs étaient capables - même si dans une moindre mesure - d'identifier et de suivre les locuteurs. Sur quels indices visuels s'appuient-ils ? Le mouvement des mains, des lèvres, le comportement de l'auditoire ? Afin de répondre à ces questions, nous présentons à 40 nouveaux participants une nouvelle base de vidéos de conversation, dont le nombre et la position relative des visages présents à l'image sont contrôlés. Ceci nous permet d'isoler plus facilement les informations liées aux différentes parties du corps des locuteurs. De plus, les stimuli que nous utilisons sont issus d'une base de vidéos librement disponible sur internet, ce qui permet (ou permettra) une comparaison plus directe de notre modèle avec ceux d'autres auteurs.

Dans un premier temps, nous décrivons en détail les stimuli utilisés ainsi que le protocole observé. Puis, nous comparons rapidement certains résultats oculométriques à ceux obtenus au chapitre précédent. Enfin, nous présentons et évaluons notre modèle de saillance audiovisuelle.

5.2 Expérience 3

5.2.1 Design Expérimental

5.2.1.1 Participants, dispositif

40 personnes ont participé à l'expérience : 28 hommes et 12 femmes, âgés entre 22 et 36 ans ($M = 26.7\,; SD = 3.5$). Les participants étaient naïfs quant au but

Figure 5.1 – **Haut** : disposition de la salle de réunion et du matériel d'enregistrement (extrait de [Aran *et al.* 2010]). **Bas** : les stimuli utilisés sont issus de la juxtaposition des prises de vue des caméras de gauche (cadre plein bleu) et droite (cadre pointillé magenta).

de l'expérience, et avaient pour consigne de regarder librement et attentivement les vidéos présentées. Tous les participants avaient une vue normale ou corrigée à la normale. Aucun n'a reporté de trouble auditif. Tous étaient titulaires d'un test d'anglais tel que le TOEFL, TOEIC, BULATS, IELTS ou équivalent. Chacun a donné son consentement éclairé à prendre part à l'expérience.
Le dispositif et l'organisation des données sont largement identiques à ceux utilisés lors de l'expérience 1 et décrits section 2.1.2.2. La seule différence réside dans l'écran utilisé pour présenter les stimuli. Comme les vidéos utilisées étaient plus larges que hautes (voir au bas de la Figure 5.1), nous avons opté pour un écran large 16 :10 LCD ViewSonic VX2268wm de 22 pouces, avec une résolution de 1280×1024 pixels et un taux de rafraichissement de 60 Hz.

5.2.1.2 Stimuli

Afin de pouvoir comparer les résultats de notre modèle avec d'autres études, nous avons utilisé une base de vidéos librement disponible sur internet, le AMI Meeting Corpus [1] [McCowan *et al.* 2005]. Il s'agit d'un corpus comprenant 100 heures d'enregistrement de réunions de travail entre quatre personnes. Nous avons choisis 3 réunions différentes (IN1008, IN1012 et IN1014), que nous avons subdivisées en 15 extraits (5 par réunions), durant entre 20 et 80 secondes ($M = 44.4$; $SD = 16.7$), ce

1. `http://www.amiproject.org/ami-scientific-portal/meeting-corpus`

qui est en moyenne plus long que pour l'expérience 2. Les réunions ont été filmées et enregistrées sous différents angles. Pour chaque frame, nous avons mis côte à côte les prises de vue des caméras latérales droite et gauche (Figure 5.1). Les stimuli ont une résolution de 1232×504 pixels (43.4×15.5 degrés) et une fréquence de 25 fps. Nous avons sélectionné des séquences où les quatre protagonistes restent assis et discutent entre eux, sans commenter ce qui est projeté. Les dialogues sont en anglais, la bande-son est monophonique et échantillonnée à 48000 Hz. La durée, le nombre de tours de parole et le temps de parole de chacun des locuteurs sont disponibles pour chaque vidéo Annexe E.

5.2.1.3 Protocole

L'enchainement des séquences et la procédure de calibration sont identiques à ceux de l'expérience 1, présentés section 2.1.2.4. Afin d'éviter tout effet d'ordre, les vidéos étaient présentées dans un ordre aléatoire. Les 20 premiers participants ont vu les 7 premières vidéos dans la condition AudioVisuelle (avec les bandes-son originales), et les 8 dernières dans la condition Visuelle (sans aucun son). L'ordre des conditions expérimentales a été contrebalancé pour les 20 participants suivants. Une pause était systématiquement proposée toutes les cinq vidéos, et les participants étaient informés qu'ils pouvaient se reposer à n'importe quel moment entre deux vidéos. Le cas échéant, une calibration était à nouveau effectuée à la reprise de l'expérience. Une expérience durait une petite demi-heure. Au final, chaque vidéo a été vue par 20 participants dans la condition Visuelle, et par 20 autres participants dans la condition AudioVisuelle.

5.2.2 Résultats

Nous comparons les conditions expérimentales Visuelle et AudioVisuelle lors de l'exploration des participants.

5.2.2.1 Dispersion et distance au centre

Pour chacune de ces métriques, nous avons mené une ANOVA à un facteur intra (la condition expérimentales). Chaque niveau a 15 items (un par stimulus). Les résultats, présentés Figure 5.2, sont en accord avec ceux obtenus pour la catégorie Visages de l'expérience 2 : avec la bande-son originale, les positions oculaires des participants sont moins dispersées ($F(1,14) = 12.8$, $p = 0.003$). Par contre, la condition expérimentale n'a pas d'effet sur la distance au centre ($F(1,14) = .86$, $p = .35$).

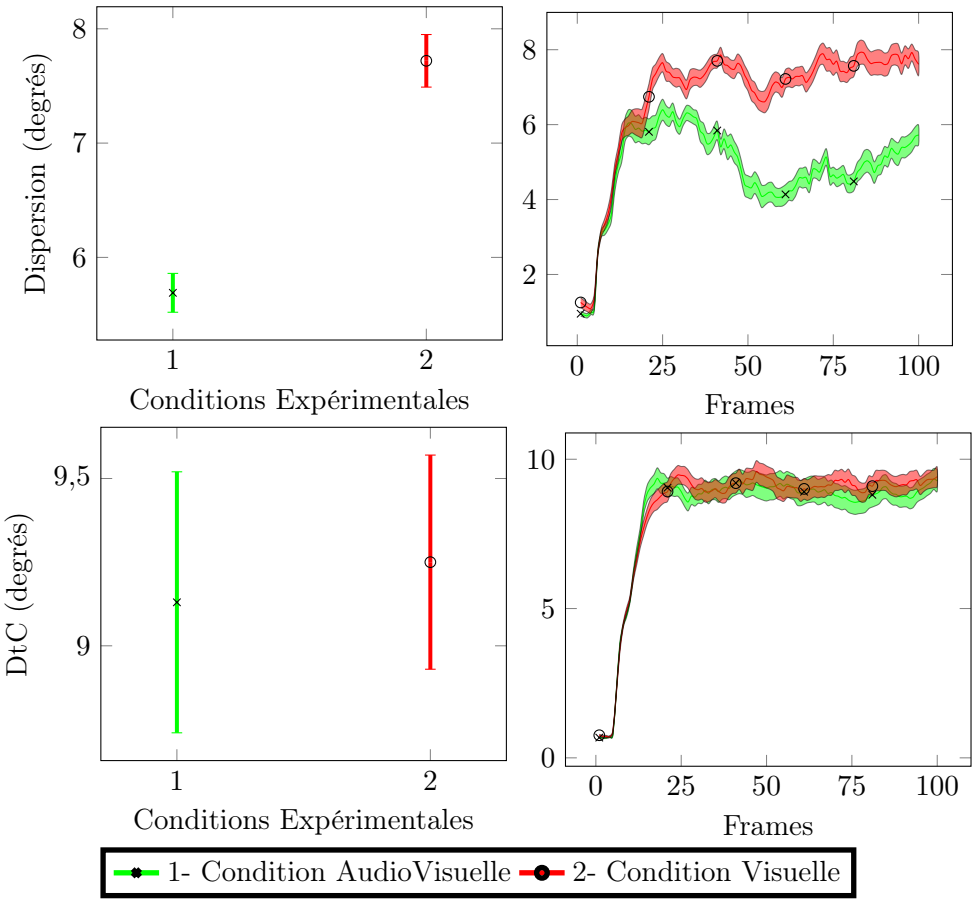


Figure 5.2 – Gauche : dispersion et distance au centre (DtC) moyennées sur l'ensemble des frames et sur l'ensemble des vidéos. **Droite :** évolutions temporelles des 100 premières frames (4 s) de la dispersion et de la distance au centre moyennées sur l'ensemble des stimuli. Les valeurs sont données en degrés angulaires, les barres d'erreur correspondent aux erreurs standards.

5.2.2.2 Amplitude de saccade et durée de fixation

Pour chacun de ces paramètres, nous avons mené une ANOVA à un facteur intra (la condition expérimentale). Chaque niveau a 40 items (un par participant). Les amplitudes de saccade ont une distribution semblable à celle constatée pour la catégorie Visages de l'expérience 2 (Figure 5.4) : un mode principal autour de 0.5°, plus prononcé pour la condition AudioVisuelle que pour la condition Visuelle, suivi d'un mode secondaire autour de 8°. Ceci se traduit par des amplitudes de saccade moyennes inférieures pour la condition AudioVisuelle, comme l'illustre la Figure 5.3 ($F(1,39) = 51$, $p < .001$).
Alors que nous n'avions pas d'effet de la condition expérimentale sur les durées de fixation pour l'expérience 2, nous mesurons ici des durées légèrement mais significativement plus courtes dans la condition Visuelle que dans la condition AudioVisuelle ($F(1,39) = 4.8$, $p = .03$). Leurs distributions demeurent inchangées, semblables à

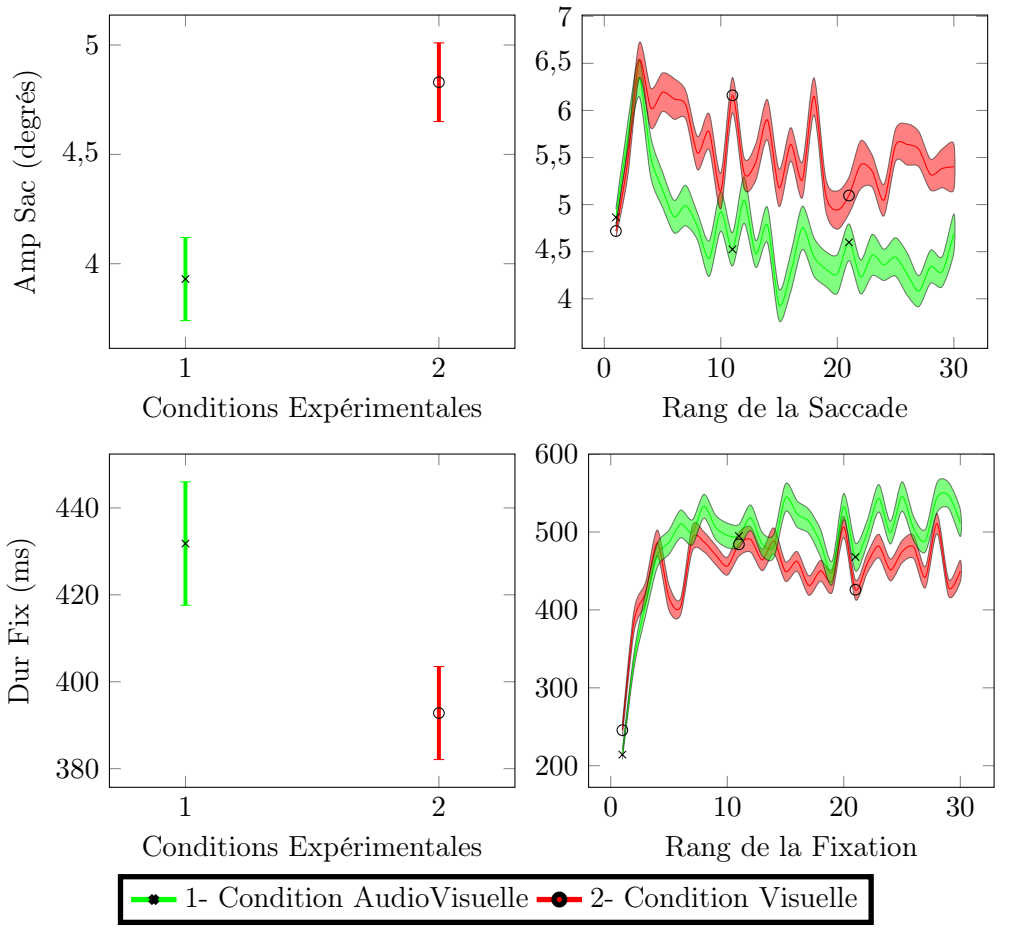


Figure 5.3 – **Gauche :** Moyenne des amplitudes de saccade et durées de fixation médianes de chaque sujet. **Droite :** amplitudes de saccade et durées de fixation en fonction de leur rang, moyennées sur chaque sujet. Les barres d'erreur correspondent aux erreurs standards.

celles obtenues aux chapitres précédents.

5.2.2.3 Proportions des fixations

Comme au chapitre précédent, nous utilisons le programme Sensarea pour définir à chaque frame des masques "visages" (2.3×2.4 degrés) correspondant aux visages des personnes présentes à l'écran. Nous avons également créé de plus grands masques "corps" (9.3×9.5 pixels) englobant aussi le torse, le visage et les mains (Figure 5.5). Afin de pouvoir distinguer les locuteurs des auditeurs, nous avons manuellement repéré pour chaque visage les périodes temporelles de parole et de silence. Pour chaque vidéo, nous calculons la moyenne des proportions de fixations atterrissant sur les locuteurs, les auditeurs, et le fond de l'image (Table 5.1). Nous avons mené une ANOVA à deux facteurs intra sur les proportions de fixations dans

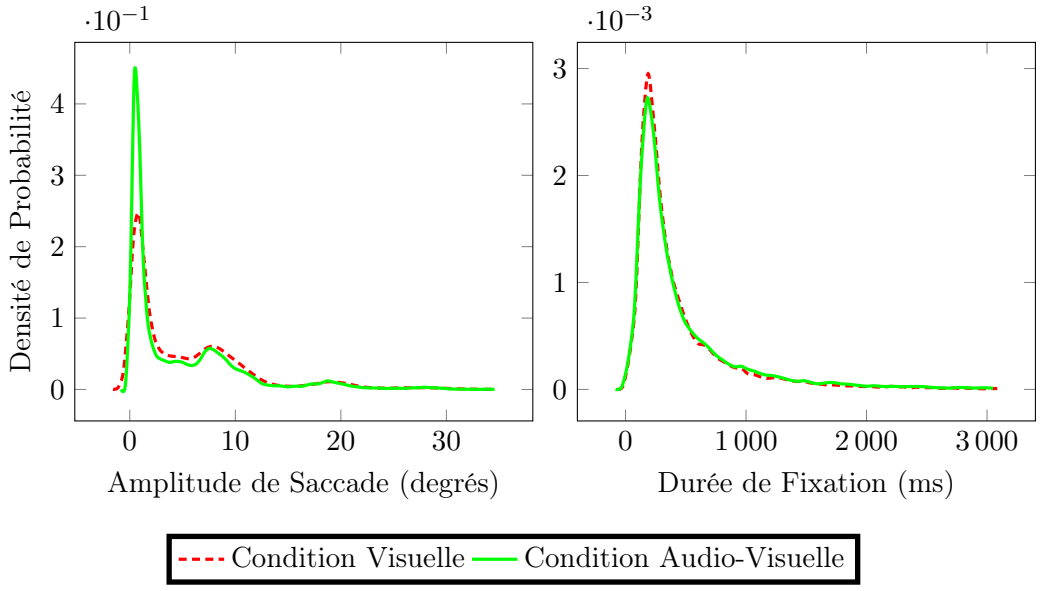


Figure 5.4 – Distributions des amplitudes de saccade et des durées de fixation médianes dans les deux conditions expérimentales. Les densités de probabilité ont été estimées sur 100 points régulièrement espacés couvrant l'ensemble des données (fonction Matlab ksdensity).

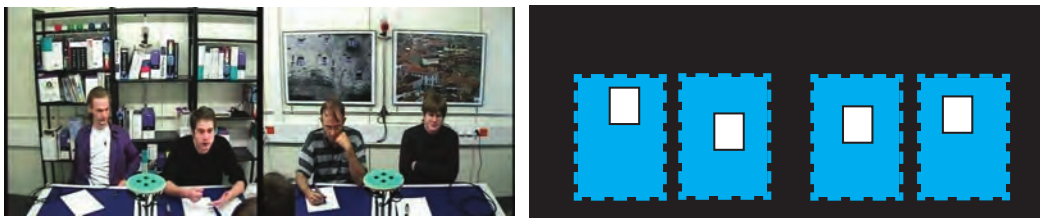


Figure 5.5 – Pour chaque frame, un masque est défini pour chaque visage (rectangles blancs) et chaque corps (rectangles bleus). Le fond de l'image correspond à la partie noire.

les masques Corps. Le premier facteur a 3 niveaux : locuteur, auditeur et fond. Le second facteur a 2 niveaux : les conditions expérimentales. Chaque niveau a 15 items (les stimuli). Il y a un effet principal des régions d'intérêt ($F(2,28) = 530$, $p < .001$), et des comparaisons a posteriori de Bonferroni indiquent que les taux de fixations sont supérieurs dans les locuteurs que dans les auditeurs, et supérieurs dans les auditeurs que dans le fond de l'image (tous les $p < .001$). Il y a également un effet principal de la condition expérimentale ($F(1,14) = 57.7$, $p < .001$) et de l'interaction ($F(2,28) = 148.6$, $p < .001$). Les locuteurs sont moins fixés dans la condition Visuelle que dans la condition AudioVisuelle, alors que c'est l'inverse pour les auditeurs (tous les $p < .001$). Par contre, les conditions expérimentales n'induisent pas de différence significative pour les taux de fixations dans le fond de l'image ($p = 1$).
Ces résultats sont identiques à ceux des taux de fixations dans les masques Visages.

Table 5.1 – Proportions des fixations par locuteur, auditeur, ou sur le fond de l'image, dans les deux conditions expérimentales. Les régions Visage et Corps sont définies Figure 5.5. La plupart du temps, il y a trois auditeurs et un locuteur, mais il arrive qu'il y ait des moments de silence ou que deux personnes parlent en même temps. Pour cette raison, la somme des fixations n'est pas exactement égale à 100%

	AudioVisuelle		**Visuelle**	
Fixations (%)	**Visage**	**Corps**	**Visage**	**Corps**
Par Locuteur	50.5 (±4)	66.3 (±5)	37.3 (±3)	53.8 (±4)
Par Auditeur	5.3 (±1)	8.3 (±1)	8.0 (±1)	13.3 (±2)
Sur le Fond		2.1 (±1)		4.0 (±2)

Ces résultats sont en accord avec ceux de la catégorie Visages de l'expérience 2 (Table 4.2) : les locuteurs se taillent la part du lion, surtout avec la bande-son originale. La plus grande différence avec ces résultats réside dans les taux de fixations par auditeur (20% avec la bande-son associée dans l'expérience 2, contre 5.3% ici). Ceci s'explique simplement par le nombre de personnes présentes par vidéo : alors que la majorité des stimuli de l'expérience 2 ne présentent que 2 individus, les stimuli de l'expérience 3 en présentent 4. Un auditeur donné a donc moins de chance d'être fixé dans cette dernière car il est concurrencé par deux autres. Nos résultats sont également proches de ceux présentés dans [Foulsham & Sanderson 2013], dont la structure des stimuli est semblable à la nôtre (Figure 4.3b et Table 4.1).

5.3 Architecture du modèle

Le modèle que nous proposons décompose chaque frame en cartes d'attributs présentées au chapitre précédent, section 4.3 : saillance statique, saillance dynamique, biais de centralité, et 4 cartes "visages" (une par personne présente à l'image). Nous rajoutons également les cartes des quatre corps correspondants (sans les visages), soit un total de 11 cartes d'attributs (Figure 5.6). Dans un premier temps, nous proposons un algorithme de *speaker diarization* capable de repérer automatiquement "qui parle quand", permettant ainsi de distinguer les locuteurs des auditeurs. Puis, nous présentons la méthode choisie pour pondérer et fusionner de manière optimale l'ensemble de ces cartes d'attributs en une carte de saillance maîtresse.

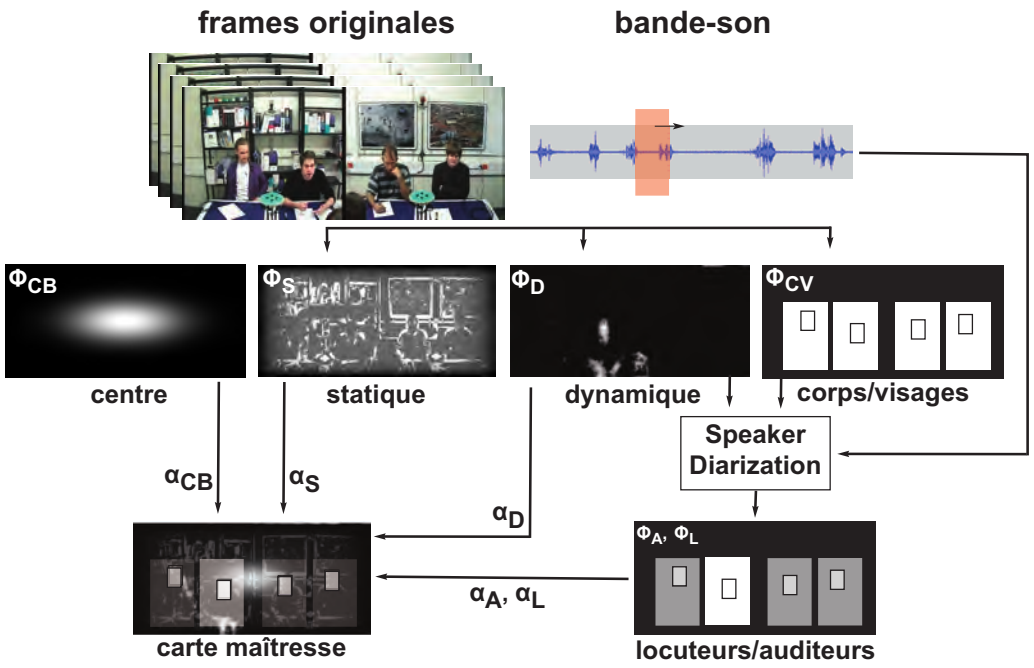


Figure 5.6 – Architecture du modèle de saillance audiovisuelle appliqué à des scènes de conversation dynamiques. En entrée du modèle, on trouve les frames originales, ainsi que la bande-son correspondante (rectangle orange). On extrait 11 cartes d'attributs visuels (Φ_i) : celles du biais de centralité CB, de la saillance statique S, de la saillance dynamique D, des 4 visages V et des 4 corps C présents à l'image. A partir des informations contenues dans la bande-son, dans la carte de saillance dynamique, et des cartes corps/visages, l'algorithme de *Speaker Diarization* permet de séparer les locuteurs des auditeurs. Enfin, toutes ces cartes sont pondérées (poids α_i) et fusionnées en une carte de saillance maîtresse.

5.3.1 Speaker Diarization

La segmentation et le regroupement en locuteurs (*speaker diarization*) sont devenus en quelques années un champ de recherche à lui seul. En effet, cette technique s'inscrit dans les méthodes d'indexation automatique des contenus multimédias, dont les bases de données chaque jour plus dantesques ont cruellement besoin. Formellement, il s'agit de détecter, et éventuellement d'identifier, les différents locuteurs s'exprimant dans un flux donné. Si le flux est purement audio, la tâche sera de repérer les périodes temporelles correspondant à chaque intervenant. Si le flux est audiovisuel, la localisation spatiale de ces derniers sera également requise. Les algorithmes de *speaker diarization* se basent généralement sur deux étapes [Anguera *et al.* 2012]. La première est la détection des segments de parole qui composent le signal sonore (*voice activity detection*). La seconde est le regroupement des différents segments de parole en locuteurs homogènes. Pour réaliser ces deux étapes, de nombreux algorithmes ont été proposés, allant de l'exploitation de l'information mutuelle entre son et image [Hershey & Movellan 1999] à la construction de modèles probabilistes, comme les chaînes de Markov cachées (HMM) ou les Deep Belief

Networks [Noulas *et al.* 2012]. La plupart de ces algorithmes extraient de l'image l'information de mouvement, et du son les coefficients cepstraux MFCC (voir plus bas section 5.3.1.2).
L'algorithme que nous proposons s'inscrit dans cette logique et s'appuie sur deux hypothèses. La première est que chaque changement de parole est séparé par un silence (aussi court soit-il). Le seconde est que le locuteur bouge plus que les personnes qui ne parlent pas. Cette dernière est liée à la communication non verbale, mode d'expression reposant notamment sur les gestes et mimiques destinés à appuyer ou moduler un discours, et richement documenté dans la littérature [McNeill 1985, Gebrekidan Gebre *et al.* 2013]

5.3.1.1 Détection des segments de parole

Pour cette étape, nous nous basons sur une méthode utilisant l'énergie et le centroïde spectral du signal sonore. Cette méthode a été initialement proposée et implémentée en langage Matlab par Theodoros Giannakopoulos [Giannakopoulos 2010]. Afin d'extraire ces attributs, le signal d'entrée est d'abord segmenté en fenêtres de 20 ms. L'énergie et le centroïde spectral sont calculés pour chacune de ces fenêtres. Soit $(x_i(n))_{n\in[1..N]}$ les échantillons audio de la i-ème fenêtre de longueur N.
Energie :

$$E(i) = \frac{1}{N}\sum_{n=1}^{N} x_i^2(n) \tag{5.1}$$

Dans un environnement sonore peu bruité, l'énergie des segments de parole est supérieure aux segments de silence, ou de bruit de fond.
Centroïde spectral : Le centroïde spectral est le centre de gravité fréquentiel d'une densité spectrale de puissance. Il est lié à la brillance d'un son, c'est-à-dire à la balance entre les fréquences graves et aigües. Soit F(k) l'amplitude de l'échantillon k de la transformée de Fourier discrète du signal

$$C(i) = \frac{\sum_{k=1}^{N} kF(k)}{\sum_{k=1}^{N} F(k)} \tag{5.2}$$

Les segments de parole sont de plus hautes fréquences que le bruit de fond.
Ces deux attributs sont calculés sur tout le signal sonore, puis une simple procédure de seuillage est appliquée pour extraire les segments de parole. Pour plus de détails, se référer à [Giannakopoulos 2010]. Afin de savoir si deux segments consécutifs ont été prononcés par deux locuteurs différents, ou s'il s'agit d'un même locuteur ayant fait une pause, nous devons attribuer à chaque segment de parole un locuteur.

5.3.1.2 Regroupement en locuteurs

Notre algorithme possède deux voies : une auditive et une visuelle. Ces voies donnent chacune un indice quantitatif de changement de locuteur, que nous combinons et seuillons afin de décider à chaque début de segment de parole s'il y a ou non changement de locuteur.

Voie Auditive

Nous nous appuyons sur une mesure de distance entre deux signaux sonores basée sur le Critère d'Information Bayésien (BIC), que nous avons déjà utilisé en modélisation statistique (équation 3.5). Cette distance, nommée ΔBIC a initialement été proposée dans [Chen & Gopalakrishnan 1998] et a depuis été utilisée avec succès pour la segmentation de locuteurs [Vajaria *et al.* 2008,Cheng *et al.* 2010]. Le principe de la méthode est le suivant :

1. Le signal de parole précédemment détecté est séparé en courts intervalles, dont sont extraits un certain nombre d'attributs bas niveau.
2. Soient deux intervalles consécutifs. La vraisemblance d'un modèle dans lequel les attributs bas niveau des deux intervalles sont issus d'une même distribution gaussienne est comparée à celle d'un modèle dans lequel ils sont issus de deux distributions gaussiennes différentes.
3. Si la vraisemblance du premier modèle (un seul mode gaussien) est supérieure à celle du second (deux modes gaussiens distincts), la même personne a prononcé les deux intervalles. Sinon, un changement de locuteur a eu lieu.

Comme attributs bas niveau, nous utilisons les coefficients cepstraux MFCC (*Mel Frequency Cepstral Coefficients*). Ces derniers, proposés en 1980 dans [Davis & Mermelstein 1980] sont les attributs les plus couramment utilisés en traitement de la parole [Noulas 2010]. Le calcul des MFCC se fait en quatre étapes. La première prend la transformée de Fourier du signal sur une fenêtre donnée, puis la seconde ordonne les puissances spectrales obtenues sur l'échelle de Mel. L'échelle de Mel est une échelle psychoacoustique de fréquences dont l'unité de mesure, basée sur la perception humaine est le mel [Stevens *et al.* 1937]. Elle a été conçue de telle sorte que la valeur en mels soit perçue par les auditeurs comme une fonction linéaire de la hauteur du son. La troisième étape prend le logarithme de la puissance dans chaque bande de fréquence, et la quatrième prend la transformée en cosinus discrète de ces logarithmes, comme s'ils constituaient un signal. Nous ne développerons pas plus dans ce manuscrit, mais le lecteur intéressé peut se reporter au cours du Professeur Barreto [2], expliquant le sens physique de ces opérations.
Nous extrayons les MFCC de nos bandes-son grâce au *HTK Hidden Markov Model Toolkit* [Young 1994]. Nous utilisons 26 coefficients, calculés sur des fenêtres de Hamming de 10 ms. Soit $\mathbf{z} = \{z_i \in \mathbf{R}^d, i = 1, ..., N\}$ le vecteur de MFCC décrivant

2. `http://www.cic.unb.br/~lamar/te073/Aulas/mfcc.pdf`

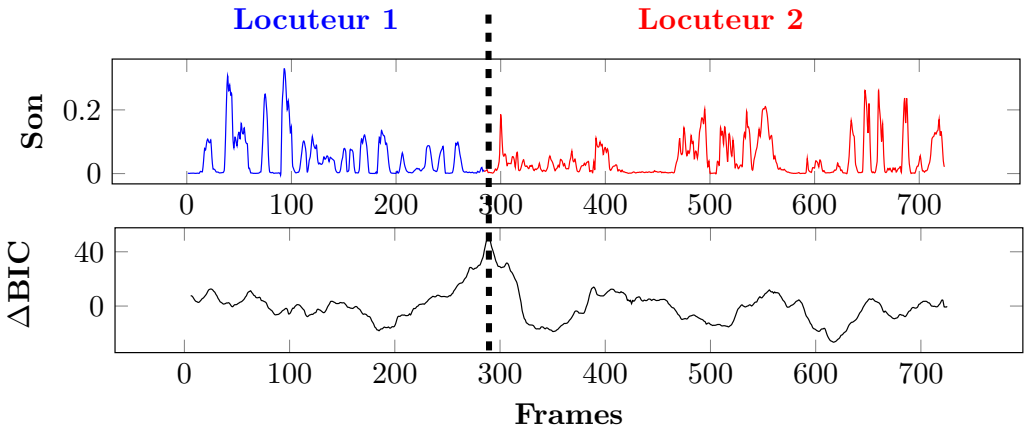


Figure 5.7 – Haut : Energie d'une bande-son dans laquelle parlent successivement le Locuteur 1 (à gauche, en bleu) et le locuteur 2 (à droite, en rouge). **Bas :** Evolution du ΔBIC correspondant. Son maximum correspond au changement de locuteur.

un signal sonore contenant N fenêtres de 10 ms. Ici $d = 26$, et le signal sonore dure N×10 ms. Une fenêtre symétrique de taille L = 200 ms est centrée sur chaque échantillon s du signal sonore. Nous testons l'hypothèse selon laquelle un changement de locuteur a lieu à l'échantillon s.
Nous comparons un modèle **H1** stipulant que les MFCC des échantillons contenus dans la fenêtre $\mathbf{w} = \{z_i \in \mathbf{R}^d, i = s - \frac{L}{2}, ..., s + \frac{L}{2}\}$ sont issus d'un processus gaussien multivarié $\mathbf{w} \sim N(\mu_w, \Sigma_w)$ avec μ_w et Σ_w la moyenne et la variance de $\mathbf{w}$; *versus* un modèle **H2** avec deux processus gaussiens différents : un pour la première moitié de la fenêtre $\mathbf{x} = \{z_{s-L/2}, ..., z_s\} \sim N(\mu_x, \Sigma_x)$, et un autre pour la seconde moitié $\mathbf{y} = \{z_s, ..., z_{s+L/2}\} \sim N(\mu_y, \Sigma_y)$.
La fenêtre d'analyse est glissée le long des segments de parole successifs, et la différence entre les BIC de ces deux modèles est calculée à chaque échantillon :

$$\begin{aligned} \Delta BIC(X, Y) &= BIC(H2, \mathbf{w}) - BIC(H1, \mathbf{w}) \\ &= \log p(X|\mu_x, \Sigma_x) + \log p(Y|\mu_y, \Sigma_y) - \log p(Z|\mu_w, \Sigma_w) - P \\ &= L \log \|\Sigma_w\| - \frac{L}{2} \log \|\Sigma_x\| - \frac{L}{2} \log \|\Sigma_y\| - P \end{aligned} \quad (5.3)$$

avec P la constante de pénalité dépendant du nombre de paramètres ($d = 26$) et de la taille de la fenêtre L. Le ΔBIC est calculé à partir de l'échantillon L/2 + 1 après le début du signal, et jusqu'à l'échantillon L/2 + 1 avant la fin du signal. Les maxima locaux de l'évolution du ΔBIC correspondent aux changements de locuteurs, comme l'illustre la Figure 5.7. Pour chaque segment de parole défini à la section précédente, nous divisons le maximum global par la moyenne des maxima locaux. Ceci donne un indice auditif qu'un changement de locuteur ait lieu dans cet intervalle.

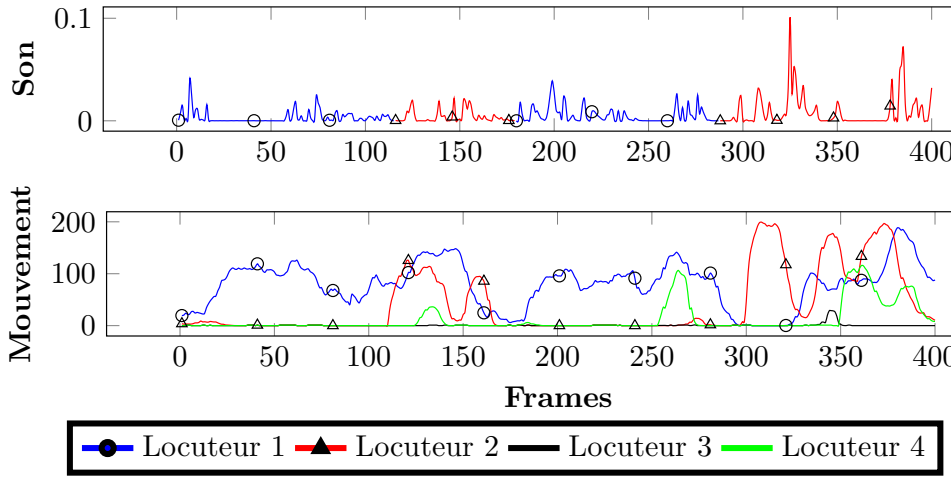


Figure 5.8 – **Haut :** Energie d'une bande-son dans laquelle parlent successivement le Locuteur 1 (en bleu, marqueurs circulaires) et le locuteur 2 (en rouge, marqueurs triangulaires). **Bas :** Evolution de l'activité des visages des 4 personnes présentes à l'image. Les périodes où les personnes sont les plus actives correspondent à leurs tours de parole. Ici, les locuteur 3 et 4 ne parlent pas.

Voie Visuelle

Nous utilisons ici l'hypothèse selon laquelle les locuteurs bougent plus que les auditeurs. Pour chaque frame, nous prenons la sortie de la voie dynamique du modèle de saillance de Marat [Marat *et al.* 2009]. Pour chaque personne présente à l'image, nous sommons la valeur des pixels contenus dans les masques "visages" ou "corps" définis Figure 5.5, et normalisons par la taille du masque. Nous avons ainsi l'évolution frame après frame de l'"activité" de chaque personne, que nous moyennons sur les périodes temporelles correspondant aux segments de parole. Plus le module de la différence de cette activité entre deux segments consécutifs est grand, plus il est probable que la personne soit passée du statut d'auditeur au statut de locuteur (ou l'inverse). La somme des différences d'activité de chaque personne fournit un indice visuel qu'un changement de locuteur ait lieu dans cet intervalle (Figure 5.8).

Pour chaque transition entre deux segments de parole, nous avons un indice auditif et un indice visuel. Une bande-son composée de M segments de parole est donc caractérisée par deux vecteurs à $M-1$ composantes I_a et I_v, contenant chacun de ces indices. Pour obtenir un vecteur d'indices bimodaux, nous centrons et réduisons I_a et I_v avant de les sommer :

$$I_{av} = \frac{I_a - mean(I_a)}{std(I_a)} + \frac{I_v - mean(I_v)}{std(I_v)} \tag{5.4}$$

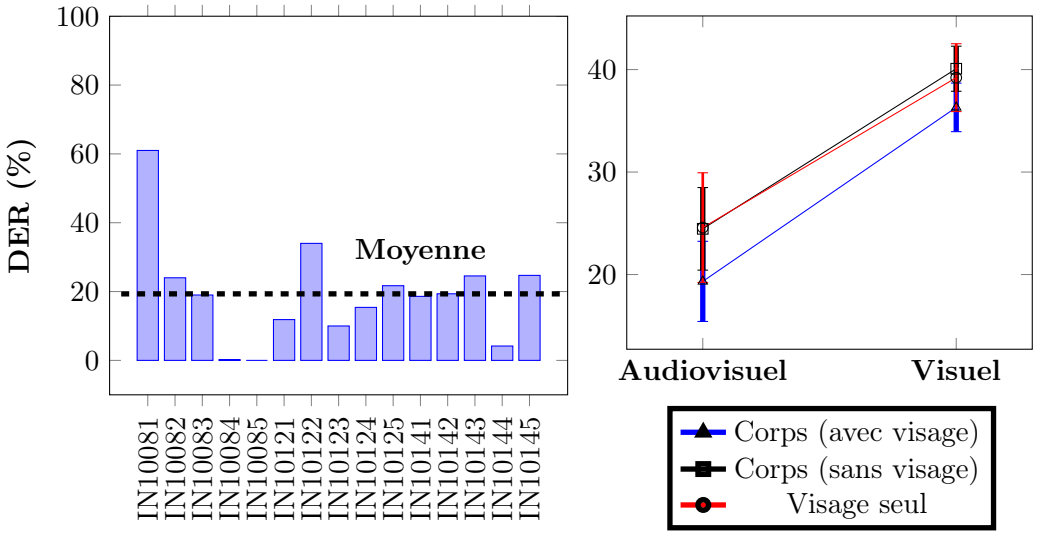


Figure 5.9 – Gauche : *Diarization Error Rate* (DER) de l'algorithme complet (Audiovisuel), l'activité étant calculée à partir des masques Corps (avec visage) définis Figure 5.5. **Droite :** DER de l'algorithme complet (Audiovisuel) ou sans sa voie auditive (Visuel), selon que l'activité est calculée à partir des masques Corps ou Visages. Plus le DER est faible, meilleur est l'algorithme.

Afin de déterminer si une transition correspond à un changement de locuteur, le vecteur I_{av} est seuillé avec un seuil déterminé empiriquement. Ici, nous utilisons un seuil de 1.6. Les segments de parole consécutifs dont l'indice de transition est inférieur au seuil sont alors fusionnés en un seul segment. Pour chacun de ces nouveaux segments, le locuteur associé est celui dont l'activité moyenne durant la période temporelle correspondante est la plus grande.

5.3.1.3 Evaluation de l'algorithme

La métrique la plus couramment utilisée pour évaluer les algorithmes de *Speaker Diarization* est le *Diarization Error Rate (DER)*, proposé par le National Institute for Standards and Technology américain [3]. Il s'agit simplement du pourcentage des frames mal classées, lors de la première étape (détection des segments de parole), ou de la seconde (regroupement en locuteurs). Pour déterminer si une frame est bien ou mal classée, nous nous basons sur une vérité terrain déterminée manuellement.
A gauche de la Figure 5.9, nous représentons les DER pour nos 15 vidéos. Nous obtenons une moyenne de 19%, ce qui est une bonne performance au regard de la littérature. En effet, les études ayant évalué leur algorithme sur la même base de vidéos que nous (AMI corpus) obtiennent des DER allant de 21% à 35% [Hung *et al.* 2008, Friedland *et al.* 2009, Gebrekidan Gebre *et al.* 2013]. Cependant, il faut garder à l'esprit que même si nous prenons la précaution d'évaluer nos algorithmes

3. `http://www.nist.gov/speech/tests/rt`

sur le même type de stimuli, comparer leurs performances reste délicat tant le DER est variable d'un stimulus à l'autre. Ceci est dû au grand nombre et à la grande variabilité des attributs sur lesquels se basent la plupart des algorithmes (nombre de locuteurs, nombre de tours de parole, durée moyenne d'un tour de parole...) [Mirghafori & Wooters 2006].
Pour évaluer notre algorithme, nous avons également fait varier deux facteurs. Le premier est le masque à l'intérieur duquel nous sommons l'activité d'une personne (Visage ou Corps). Le second est l'efficacité de la voie auditive dans le regroupement en locuteurs. Ce second facteur nous permet de comparer notre algorithme audiovisuel à un algorithme purement visuel, faisant fi de la bande-son. Dans cet algorithme visuel, le statut de locuteur est simplement attribué à la personne présentant la plus forte activité visuelle à une frame donnée. L'activité visuelle est définie comme la quantité de mouvement de la carte dynamique de Marat *et al.* sur toute la surface du masque considéré, et normalisée par la taille de ce dernier. Les résultats, à droite de la Figure 5.9 indiquent que la voie auditive est nécessaire au bon fonctionnement de l'algorithme : sans cette dernière, les performances sont nettement dégradées (DER autour de 40%). Nous avons mené une ANOVA à deux facteurs intra. Le premier est le type de masque et : visage seul, corps + visage, corps seul. Le second est le type d'algorithme : visuel ou audiovisuel. Chaque niveau possède 15 items (le nombre de vidéos). L'algorithme audiovisuel est plus performant que l'algorithme visuel ($F(1,14) = 24.4$, $p < .001$), mais nous n'avons pas trouvé d'effet du type de masque ($F(2,28) = 2.5$, $p = .1$), ni de l'interaction ($F(2,28) = .17$, $p = .84$).

Cet algorithme a également été évalué sur les vidéos de la catégorie Visages de l'expérience 2. Nous ne présentons pas les résultats ici, mais ils sont présentés en détail dans [Coutrot & Guyader 2014a]. Il sont un peu moins bons que ceux que nous venons de présenter, car ces vidéos sont plus bruitées (personnes souvent en mouvement dans un environnement complexe), mais restent toutefois au niveau de la littérature.

5.3.2 Fusion

Nous proposons une fusion originale des différentes cartes d'attributs en une carte de saillance maîtresse mettant en évidence les régions les plus susceptibles d'attirer l'attention. Au lieu de faire une simple moyenne avec toutes les cartes, ou de les pondérer en fonction de leur propriétés physiques (maxima, distribution...), nous estimons leur poids par une modélisation statistique de type Lasso, décrite section 3.3.1.2. Pour chaque vidéo, nous moyennons le poids de chaque attribut sur toutes les frames. Puis, nous prenons la valeur moyenne de ces poids sur l'ensemble des vidéos, et les affectons aux cartes correspondantes. Les poids utilisés dans les modèles évalués ci-dessous sont ceux calculés à partir des positions oculaires enregistrées dans la condition AudioVisuelle. Les cartes "Corps" occupant une aire importante

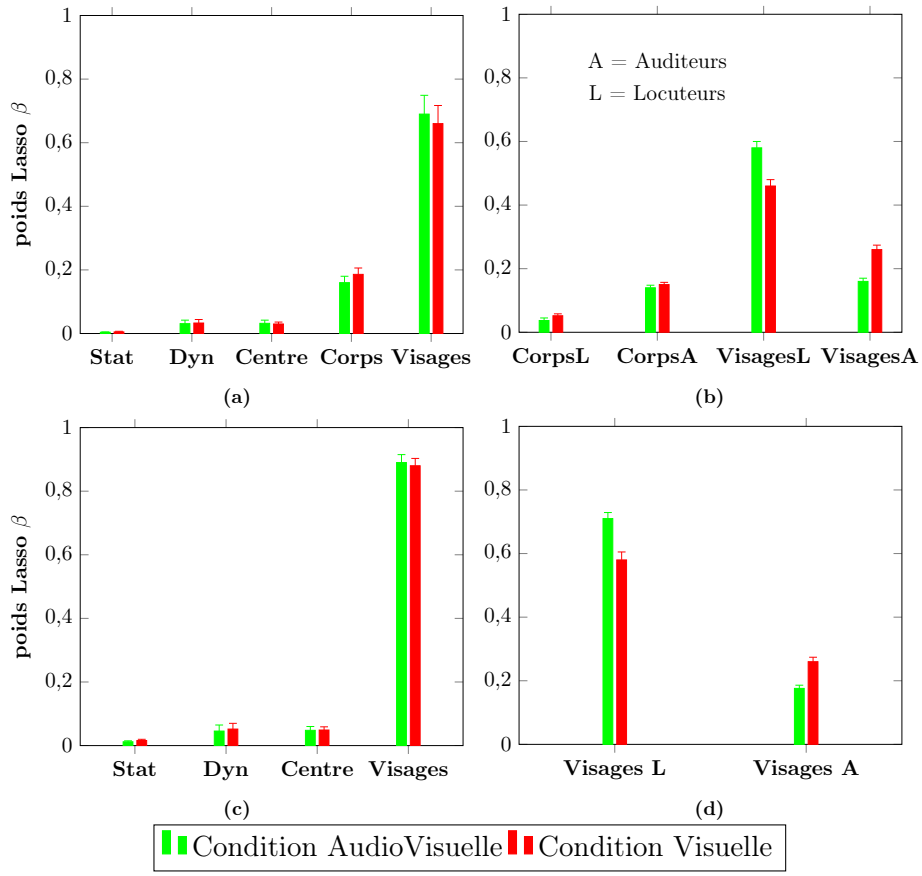


Figure 5.10 – **(a)** Valeurs moyennes des poids Lasso des attributs saillance statique, saillance dynamique, biais de centralité, corps et visages. ($\sum_{i=1}^{5} \beta_i = 1$). **(b)** Contributions des visages et des corps (sans visage) des locuteurs (L) et auditeurs (A) aux attributs "Visages" et "Corps" de la figure de gauche. **(c et d)** Idem sans l'attribut corps.
Les poids sont moyennés sur toutes les frames de chaque vidéo, puis sur l'ensemble des vidéos. Les barres d'erreurs correspondent aux erreurs standards.

des vidéos et pouvant biaiser la contribution d'autres attributs, nous avons estimé les poids de deux modèles : l'un prenant en compte les corps et les visages de chaque protagoniste (Figures 5.10a et 5.10b), et l'autre ne prenant en compte que les visages (Figures 5.10c et 5.10d).

5.4 Evaluation du modèle

5.4.1 Différents attributs, différentes fusions

Nous évaluons notre modèle en comparant les régions prédites comme étant saillantes avec les régions effectivement regardées par les participants de l'expérience 3. Nous utilisons la divergence de Kullback-Leibler (DKL, équation 1.3) et le

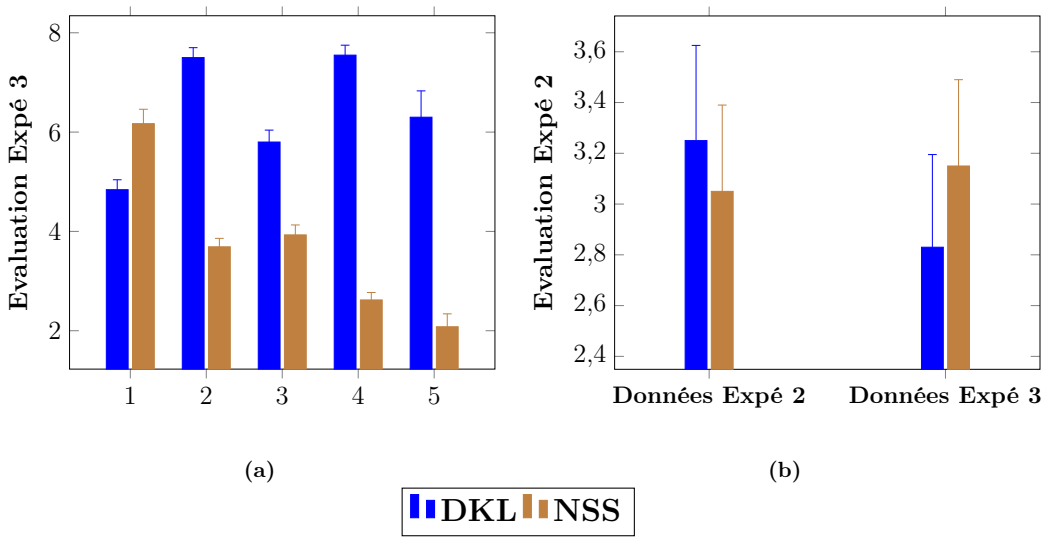


Figure 5.11 – **(a)** NSS et DKL des différents modèles décrits dans cette section, appliqués aux vidéos de l'expérience 3. Les poids des attributs ont été estimés par Lasso à partir des positions oculaires enregistrées dans la condition AudioVisuelle de l'expérience 3.
(b) NSS et DKL du modèle 1 appliqué aux vidéos de la catégorie Visages de l'expérience 2. Les poids des attributs ont été estimés par Lasso à partir des positions oculaires enregistrées dans la condition AudioVisuelle de l'expérience 3 (droite) ainsi que dans la condition Originale de l'expérience 2 (gauche). Pour rappel, un modèle est d'autant meilleur que le NSS est grand et la DKL petite. Les barres d'erreurs correspondent aux erreurs standards.

Normalized Scanpath Saliency (NSS, équation 1.4). Figure 5.11a, nous comparons les résultats calculés à partir des positions oculaires enregistrées dans la condition AudioVisuelle pour les modèles utilisant les attributs et modes de fusion suivants :

1. fusion Lasso de saillance statique, saillance dynamique, biais de centralité, visages des locuteurs et des auditeurs (Figures 5.10c et 5.10d)
2. fusion Lasso de saillance statique, saillance dynamique, biais de centralité, visages et corps des locuteurs et des auditeurs (Figures 5.10a et 5.10b)
3. fusion Lasso de saillance statique, saillance dynamique, biais de centralité, visages poids égaux et constants (Figure 5.10c)
4. simple moyenne de saillance statique, saillance dynamique, biais de centralité et visages.
5. modèle bas niveau (uniquement saillance statique et dynamique) et fusion proposés dans [Marat *et al.* 2009].

Pour ne pas évaluer le modèle avec les mêmes positions oculaires que celles qui nous ont servi à le construire, nous utilisons la méthode du *leave-one-out*. Plus précisément, les poids des attributs utilisés pour construire le modèle d'une vidéo donnée sont issus de la moyenne des poids de toutes les vidéos, sauf de celle traitée.

Nous observons que NSS et DKL donne des résultats concordants : lorsque le NSS d'un modèle est grand, sa DKL est petite.

Nous avons mené deux ANOVA à un facteur intra (les différents modèles) sur les valeurs moyennes de NSS et de DKL. Il existe bien un effet principal des modèles sur les NSS ($F(4,56) = 453.7$, $p < .001$), ainsi que sur les DKL ($F(4,56) = 78.9$, $p < .001$). Le meilleur modèle est sans conteste le premier, qui distingue les visages des locuteurs de ceux des auditeurs (son NSS est le plus grand et sa DKL la plus petite, tous les $p < .001$).
Contrairement à ce que nous attendions, le second modèle, qui prend en plus en compte le corps des intervenants (torse + mains), présente de bien moins bonnes performances, comparables en NSS à celles du modèle donnant un poids égal et constant à tous les visages ($p = .20$), et en DKL à celles du modèle prenant simplement la moyenne des attributs ($p = 1$). Ceci est sans doute dû à la grande surface des corps comparée à leur attractivité. Il pourrait être intéressant de quantifier plus précisément la contribution de plus petites parties du corps, comme les mains ou le torse.
Ne pas distinguer le visage des auditeurs de celui des locuteurs (modèles 3 et 4) dégrade également les prédictions du modèle, et ne pas prendre du tout en compte les visages (modèle 5) conduit, comme nous l'attendions, aux plus mauvaises prédictions. A part pour la DKL des modèles 3 et 5 ($p = .15$), toutes les autres différences présentées Figure 5.11a sont significatives (tous les $p < .001$).

5.4.2 Généralisabilité des poids estimés

Nous regardons à quel point les poids estimés par NSS grâce aux positions oculaires enregistrées sur une base de vidéos de conversation donnée peuvent se généraliser à une autre base de vidéos. Nous avons appliqué le modèle 1 aux vidéos de la catégorie Visages de l'expérience 2 (scènes de conversation dans un environnement complexe et dynamique, voir Annexe B). A gauche de la Figure 5.11b sont affichées les valeurs de NSS et DKL calculées avec les poids des attributs estimés par NSS à partir des positions oculaires enregistrées dans la condition Originale de l'expérience 2 (poids consignés Figure 4.8). A droite de la Figure 5.11b sont affichées les valeurs de NSS et DKL calculés avec les poids des attributs estimés par Lasso à partir des positions oculaires enregistrées dans la condition AudioVisuelle de l'expérience 3 (poids consignés Figure 5.10).
Nous constatons que comme pour l'algorithme de *Speaker Diarization*, notre modèle est moins performant sur les vidéos de l'expérience 2, ce qui est logique au vu de leur caractère plus bruité (personnes souvent en mouvement dans un environnement complexe). Nous constatons également que les performances des modèles bâtis à partir des poids estimés grâce aux positions oculaires des expériences 2 et 3 sont comparables. Ceci indique la bonne généralisabilité de la fusion Lasso présentée dans ce chapitre.

Dans ce chapitre, nous avons présenté une base de vidéos de conversation

permettant une quantification précise et réplicable des différents paramètres susceptibles d'attirer l'attention. Nous avons analysé en détail les mouvements oculaires de 40 participants les ayant regardées avec et sans leurs bandes-son originales. Enfin, nous nous sommes servis de ces résultats pour construire un modèle d'attention visuelle appliqué aux scènes de conversation.
Les modèles de saillance classiques ne prennent pas en compte la dimension sociale de l'exploration visuelle, et ont de fait de très mauvaises performances pour ce type de scène. Afin de les améliorer, certains auteurs ont proposé de détecter les visages et d'en augmenter la saillance. Ici, nous sommes allés plus loin en distinguant le visage des locuteurs de celui des auditeurs au moyen d'un algorithme de *speaker diarization*. Nous nous servons des poids calculés par modélisation statistique (NSS) afin de résoudre le problème de fusion des différentes cartes d'attributs.
Nous montrons que ceci permet de considérablement améliorer la prédiction des zones saillantes dans des scènes de conversation. De plus, cette méthode est robuste, car les poids estimés pour une base de vidéos de conversation donnée parviennent très bien à modéliser la saillance d'une autre base différente à bien des égards.

Chapitre 6

Synthèse et perspectives

Pour clore ce manuscrit, nous rappelons les éléments saillants de ce travail de thèse. Puis, nous suggérons quelques idées pour poursuivre cette quête vers la compréhension et la modélisation des mécanismes de l'intégration audiovisuelle.

6.1 Synthèse des principaux résultats

Dans cette thèse, nous nous sommes livrés à une analyse et une quantification fine de l'influence de différents attributs audiovisuels sur l'exploration visuelle de scènes naturelles dynamiques. Nos résultats s'appuient sur 3 expériences au cours desquelles les mouvements oculaires de 148 participants ont été enregistrés sur un total de plus de 75 400 frames (125 vidéos), dans 5 conditions expérimentales différentes. L'ensemble de ces données est librement disponible sur internet [1].

La première expérience oculométrique nous a permis de démontrer que si la façon dont nous explorons une scène dépend avant tout de son contenu visuel, l'absence de son modifie certains paramètres des mouvements oculaires. Sans le son, la dispersion entre les positions oculaires des différents observateurs est plus grande, et les amplitudes de saccade plus petites. De plus, deux observateurs dans la même condition expérimentale (Visuelle ou AudioVisuelle) regardent davantage les mêmes régions que deux observateurs dans deux conditions différentes [Coutrot *et al.* 2012]. Grâce à l'implémentation d'un modèle de saillance sonore, nous avons pu comparer les mouvements oculaires moyens à ceux effectués juste après les événements auditifs les plus marquants. De précédentes études ont suggéré que saillance sonore et visuelle se combinent linéairement. L'absence d'effet de la proximité de ces pics de saillance sonore suggère au contraire que les liens entre ces deux modalités sont de nature plus complexe [Coutrot *et al.* 2014].

L'intégration audiovisuelle peut opérer de manière radicalement différente en fonction du contenu auditif et visuel présent dans la scène considérée. Nous avons donc mené une deuxième expérience oculométrique, dont les stimuli étaient classés en 4 catégories visuelles (Visages, Paysages, Un ou Plusieurs Objets en Mouvement),

1. `http://www.gipsa-lab.fr/~antoine.coutrot/`

et présentés selon 4 conditions expérimentales (avec leur bande-son originale ou avec la bande-son d'autres vidéos). Nous avons montré que l'exploration visuelle est drastiquement influencée par la catégorie visuelle. Par contre, les conditions expérimentales n'ont d'effet qu'au sein de la catégorie Visages, présentant des scènes de conversations [Coutrot & Guyader 2013a, Coutrot & Guyader 2013b]. Dans cette catégorie visuelle, la présence du signal de parole associé augmente la cohérence entre les scanpaths. Les observateurs effectuent davantage de petites saccades sur le visage des locuteurs, et changent de cible au rythme des tours de parole [Coutrot & Guyader 2014b]. Par contre, nous n'avons trouvé aucune différence entre les conditions expérimentales incongruentes. Une modélisation statistique nous a permis de quantifier l'importance relative de différents attributs visuels (saillance statique, saillance dynamique, biais de centralité, carte uniforme, visages) pour expliquer les positions oculaires enregistrées. Nous montrons que la saillance bas niveau présente un poids très faible comparé à celui des visages, et tout particulièrement du visage des locuteurs.

Forts de ce constat, nous avons mené une troisième expérience oculométrique avec une nouvelle base de vidéos de conversation, librement disponible sur internet [2]. Dans ces vidéos, le nombre et la position relative des visages présents à l'image sont contrôlés, et ces derniers évoluent dans un lieu clos. Ceci nous permet d'isoler les informations liées aux différentes parties du corps des locuteurs, et d'estimer plus facilement leur contribution dans la distribution des positions oculaires. Cette expérience nous a permis de proposer un modèle de saillance audiovisuelle adapté aux scènes de conversation [Coutrot & Guyader 2014a]. Ce modèle inclut un algorithme de segmentation des locuteurs permettant de repérer automatiquement le visage de ces derniers afin d'en rehausser la saillance. Ce modèle présente de bien meilleures performances que ceux attribuant un poids constant et égal à tous les visages présents dans la scène.

L'ensemble de ces résultats a permis de lever un bout du voile sur certains processus audiovisuels lors de l'exploration de scènes naturelles dynamiques. Cependant, la nature du lien entre saillance visuelle et saillance sonore reste mal comprise. Trouver un cadre général expliquant comment ces deux modalités interagissent pour guider l'attention demeure une entreprise ambitieuse et fascinante.

2. `http://www.amiproject.org/ami-scientific-portal/meeting-corpus`

6.2 Saillance audiovisuelle ≠ saillance sonore + saillance visuelle

Derrière ce titre tout droit sortie de la *Gestaltpsychologie*[3] du début du XX$^{\text{ème}}$ siècle, se cache une des questions restant en suspens à la fin de ce manuscrit. Il s'agit de l'absence d'effet du son sur l'exploration de scènes contenant plusieurs objets en mouvement (catégorie POM de l'expérience 2). En effet, nous pouvions penser que lorsque plusieurs objets particulièrement saillants (puisqu'en mouvement) sont en compétition, celui dont la saillance sonore est la plus forte guide davantage le regard des observateurs avec la bande-son originale qu'avec des bandes-son incongruentes. Une interprétation possible de cette absence de résultat est la suivante. Si l'on considère que l'effet du son sur les mouvements oculaires est ponctuel (par exemple fort durant un court instant T et négligeable le reste du temps), une analyse globale telle que nous l'avons menée risque de le noyer, et donc de passer à côté. Mais me direz-vous, n'est-ce pas justement pour cela que vous avez comparé les mouvements oculaires moyens avec ceux effectués juste après les événements sonores les plus saillants ? Et vous n'avez pourtant rien trouvé ? Oui, mais rétrospectivement, la façon dont nous avons défini nos pics de saillance sonore pourrait être inefficace. Ils ne sont calculés qu'à partir de l'information sonore (énergie, amplitude, fréquence). Or, comme nous l'avons constaté au chapitre 4 (à l'instar de Nahorna *et al.* et de Michel Chion, voir section 4.4.3), l'intégration audiovisuelle n'opère que si il y a liage entre les deux modalités. Aussi, plutôt que de calculer des pics de saillance sonore, il serait intéressant d'apprendre à déterminer quand un objet visuel donné est susceptible d'être lié à un objet sonore concomitant. Mais sur quels facteurs s'appuyer parmi les innombrables attributs susceptibles d'être corrélés, et donc de lier ces signaux de dimensions si différentes ?

Une mesure adaptée à la detection des corrélations audiovisuelles

Une approche intéressante est celle adoptée notamment par quelques auteurs de la communauté *computer vision* : l'analyse canonique des corrélations (*Canonical Correlation Analysis*, CCA) [Hotelling 1936]. Cette technique permet de mettre en évidence des proximités entre deux ensembles de données de dimension différente. Une démarche fréquemment adoptée est de mettre dans deux sacs différents de multiples attributs appartenant aux deux modalités : mouvement, contraste, intensité... pour le signal visuel et MFCC, *zero-crossing*, énergie, hauteur... pour le signal sonore. La CCA va alors trouver les combinaisons linéaires entre les attributs des deux sacs présentant un maximum de corrélation l'un avec l'autre. Cette méthode a par exemple été utilisée pour synchroniser un signal de parole avec les lèvres du locuteur [Slaney & Covell 2000], ou pour localiser une source sonore dans

3. Voir glossaire. Un des principes phare de cette théorie est que *le tout est différent de la somme de ses parties.*

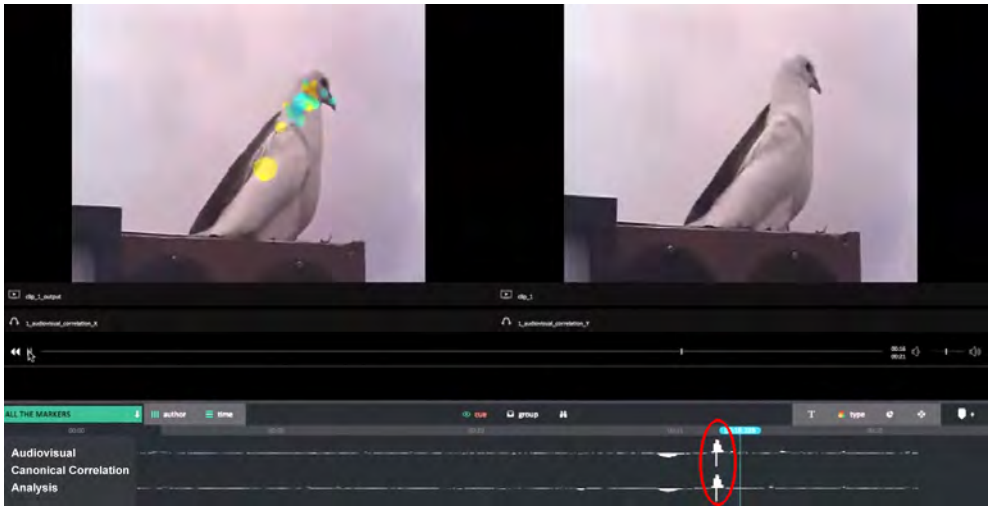


Figure 6.1 – capture d'écran du logiciel en cours de développement Rekall permettant la visualisation simultanée de plusieurs vidéos et de plusieurs courbes. En haut : frame issue de la vidéo 1 de l'expérience 1, avec et sans les positions oculaires. En bas : analyse canonique des corrélations entre les informations visuelle et sonore. La barre verticale bleue marque la progression de la vidéo sur la courbe. La vidéo complète est disponible sur internet : `https://www.youtube.com/watch?v=6L0zOf7_lH0`

une vidéo [Kidron *et al.* 2007].

Pour s'assurer de la pertinence de cette mesure, il peut s'avérer pratique de mettre directement en regard son évolution temporelle avec celle de la vidéo traitée.
La Figure 6.1 est une capture d'écran du logiciel Rekall (actuellement en cours de développement), extrêmement utile pour visualiser les données qui nous intéressent [Bardiot *et al.* 2014]. Les deux images de colombes en haut correspondent à une frame issue de la vidéo 1 de l'expérience 1 (à droite frame originale, à gauche avec les positions oculaires incrustées grâce au logiciel CARPE [4]). La courbe inférieure correspond à l'évolution de la CCA entre différents attributs visuels et sonores. On peut voir que la courbe est assez plate, jusqu'au pic repéré par l'ellipse rouge. Ce moment correspond à la sortie brusque de la colombe, parfaitement corrélée avec un bruit saillant. Ces analyses préliminaires ont été menées au mois de juin 2014 dans le cadre du projet Auracle de l'*International Summer Workshop on Multimodal User Interfaces* [Frisson *et al.* 2014].

Une bonne idée pourrait donc être de refaire les analyses présentées dans ce manuscrit au voisinage de ce type de pic. D'un point de vue modèle, cela pourrait se traduire par une mise en valeur du principal objet saillant durant les quelques frames suivant les maxima de la CCA.
Repérer les corrélations entre facteurs de bas niveau susceptibles de lier un objet visuel à un objet sonore constituerait un grand pas en avant dans la modélisation

4. `http://thediemproject.wordpress.com/software/`

Figure 6.2 – Exemple de fonctionnement d'un modèle de saillance audiovisuelle bayésien. En entrée (à gauche), une scène audiovisuelle présentant un bateau, des oiseaux, et le cri de ces derniers. Si le modèle a par le passé appris à associer cris et oiseaux, il pourra les mettre en valeur au détriment du bateau. Les cartes de saillance (sans les ellipses claires) sont issues du modèle de Marat *et al.*

de la saillance audiovisuelle. Mais il est également essentiel de considérer les correspondances de haut niveau.

Des correspondances audiovisuelles de haut niveau

Nous apprenons, par expérience, à lier entre eux des événements bimodaux récurrents [5] [Shams & Kim 2010, Spence 2011]. Ainsi, sur le long terme, nous avons appris à associer plus volontiers un son de parole à un humain qu'à tout autre objet visuel. A plus court terme, par exemple celui du contexte narratif d'un film, les liens se tissant progressivement entre les différents objets d'une scène influencent fortement la perception de cette dernière. Autrement dit, si un son a été associé à un objet par le passé, il y a plus de chance qu'il le soit également dans le futur, et leur intégration en sera facilitée.

Le formalisme bayésien est une cadre approprié à la modélisation de ce type de phénomène [Ernst & Banks 2002]. En deux mots, ce paradigme stipule que la connaissance que l'on a d'une variable donnée est une combinaison de connaissances préalables et de nouvelles observations. ll pourrait être intéressant de se servir de ce cadre théorique pour construire un nouveau modèle de saillance audiovisuelle mettant davantage en valeur les régions précédemment associées avec un son lorsque ce dernier se fait entendre (Figure 6.2). Ce projet, monté conjointement avec Bob Carlyon de l'université de Cambridge (Royaume-Uni), a fait l'objet d'une demande

5. Par exemple, nous associons plus souvent des sons aigus avec des objets petits, brillants, et situés en hauteur [Evans & Treisman 2010].

de bourse post-doctorale *Newton International Fellowship*.

Une fois ces modèles audiovisuels construits (ce qui promet quelques belles heures de travail), il convient de trouver un moyen de comparer leurs performances aux nombreux modèles unimodaux proposés dans la littérature.

Evaluation des modèle de saillance : quelle vérité terrain ?

Afin de rendre équitable la comparaison des nombreux modèles de saillance visuelle présentés dans la littérature, certains auteurs ont proposé de les évaluer sur une vérité terrain commune [Judd *et al.* 2012, Borji *et al.* 2012]. Il s'agit d'une tâche délicate tant le nombre de paramètres à contrôler est élevé : les performances d'un modèle peuvent radicalement changer en fonction des métriques utilisées [Riche *et al.* 2013b], de la taille des zones saillantes détectées ou de la prise en compte du biais centré [Riche *et al.* 2013a]. Comme nous l'avons vu en introduction (section 1.2.3.3), les modèles de saillance sont le plus souvent évalués en comparant leurs prédictions avec les régions effectivement regardées par des observateurs. Seulement là encore, le son n'est jamais pris en compte lors de l'enregistrement des positions oculaires ce qui, nous l'avons vu dans cette thèse, est susceptible fausser ces bancs d'évaluation.

Un projet en cours de réalisation en collaboration avec Nicolas Riche de l'Université de Mons, en Belgique, est d'évaluer les principaux modèles de saillance de la littérature sur une base de vérité terrain constituée de positions oculaires enregistrées sur des vidéos vues avec et sans leurs bandes-son. Ceci permettra de vérifier si les classements changent entre les deux conditions expérimentales.

Bibliographie

[Alsius & Soto-Faraco 2011] Agnès Alsius et Salvador Soto-Faraco. *Searching for audiovisual correspondence in multiple speaker scenarios.* Experimental Brain Research, vol. 213, pages 175–183, 2011. (Cité en pages 95 et 96.)

[Altman 1980] Rick Altman. *Moving Lips : Cinema as Ventriloquism.* Yale French Studies, vol. 60, pages 67–79, 1980. (Cité en page 93.)

[Anguera *et al.* 2012] Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland et Oriol Vinyals. *Speaker Diarization : A Review of Recent Research.* IEEE Transaction on Audio, Speech and Language Processing, vol. 20, no. 2, pages 356–370, 2012. (Cité en page 121.)

[Antes 1974] James R Antes. *The time course of picture viewing.* Journal of experimental psychology, vol. 103, no. 1, pages 62–70, 1974. (Cité en pages 12 et 86.)

[Aran *et al.* 2010] Oya Aran, Hayley Hung et Daniel Gatica-Perez. *A Multimodal Corpus for Studying Dominance in Small Group Conversations.* In LREC workshop on Multimodal Corpora : Advances in Capturing, Coding and Analyzing Multimodality, Malta, 2010. (Cité en page 115.)

[Arndt & Colonius 2003] Petra A. Arndt et Hans Colonius. *Two stages in crossmodal saccadic integration : evidence from a visual-auditory focused attention task.* Experimental Brain Research, vol. 150, pages 417–426, 2003. (Cité en pages 26, 27 et 48.)

[Baddeley & Tatler 2006] Roland J Baddeley et Benjamin W Tatler. *High frequency edges (but not contrast) predict where we fixate : A Bayesian system identification analysis.* Vision research, vol. 46, no. 18, pages 2824–2833, 2006. (Cité en page 76.)

[Bahill *et al.* 1975] Terry Bahill, Deborah Adler et Lawrence Stark. *Most naturally occurring human saccades have magnitudes of 15 degrees or less.* Investigative Ophthalmology, vol. 14, no. 6, pages 468–469, 1975. (Cité en page 108.)

[Bailly *et al.* 2010] Gérard Bailly, Stephan Raidt et Frédéric Elisei. *Gaze, conversational agents and face-to-face communication.* Speech Communication, vol. 52, pages 598–612, 2010. (Cité en page 91.)

[Bailly *et al.* 2012] Gérard Bailly, Pascal Perrier et Eric Vatikiotis-Bateson. Audiovisual Speech Processing. Cambridge University Press, Cambridge, UK, 2012. (Cité en pages 91 et 97.)

[Bardiot *et al.* 2014] Clarisse Bardiot, Guillaume Jacquemin, Guillaume Marais et Thierry Coduys. *REKALL : un environnement open-source pour documenter, analyser les processus de création et simplifier la reprise des œuvres.* In Actes des Journées d'Informatique Musicale, Bourges, France, 2014. (Cité en page 136.)

[Belopolsky & Theeuwes 2009] Artem V. Belopolsky et Jan Theeuwes. *When Are Attention and Saccade Preparation Dissociated ?* Psychological Science, vol. 20, no. 11, pages 1340–1347, 2009. (Cité en page 5.)

[Bertelson & de Gelder 2004] P. Bertelson et B. de Gelder. *The psychology of multimodal perception.* In C Spence et J Driver, editeurs, Crossmodal space and crossmodal attention, pages 141–177. Oxford University Press, Oxford, UK, 2004. (Cité en pages 93 et 94.)

[Berthommier 2004] F Berthommier. *A phonetically neutral model of the low-level audiovisual interaction.* Speech Communication, vol. 44, no. 1-4, pages 31–41, 2004. (Cité en page 110.)

[Bertolino 2012] Pascal Bertolino. *Sensarea : an Authoring Tool to Create Accurate Clickable Videos.* In 10th Workshop on Content-Based Multimedia Indexing, pages 1–4, Annecy, France, 2012. (Cité en page 100.)

[Bindemann *et al.* 2007] Markus Bindemann, A Mike Burton, Stephen R H Langton, Stefan R Schweinberger et Martin J Doherty. *The control of attention to faces.* Journal of Vision, vol. 7, no. 10, pages 1–8, 2007. (Cité en page 91.)

[Birmingham & Kingstone 2009] E. Birmingham et A. Kingstone. *Saliency does not account for fixations to eyes within social scenes.* Vision Research, vol. 49, pages 2992–3000, 2009. (Cité en pages 92, 108 et 113.)

[Birmingham *et al.* 2009] E. Birmingham, W. F. Bischof et A. Kingstone. *Human social attention.* Annals of the New York Academy of Sciences, vol. 1156, no. 1, pages 118–140, 2009. (Cité en pages 90 et 91.)

[Boccignone *et al.* 2005] Giuseppe Boccignone, Angelo Chianese, Vincenzo Moscato et Antonio Picariello. *Foveated Shot Detection for Video Segmentation.* IEEE Transaction on Circuits and Systems for Video Technology, vol. 15, no. 3, pages 1–13, 2005. (Cité en pages 9 et 46.)

[Bolivar *et al.* 1994] Valerie J. Bolivar, Annabel J. Cohen et John C. Fentress. *Semantic and Formal Congruency in Music and Motion Pictures : Effects on the Interpretation of Visual Action.* Psychomusicology, vol. 13, pages 28–59, 1994. (Cité en page 58.)

[Boltz *et al.* 2009] Marilyn G. Boltz, Brittany Ebendorf et Benjamin Field. *Audiovisual Interactions : The Impact of Visual Information on Music Perception and Memory.* Music Perception, vol. 27, no. 1, pages 43–59, 2009. (Cité en page 58.)

[Boltz 2004] Marilyn G. Boltz. *The cognitive processing of film and musical soundtracks.* Memory & Cognition, vol. 32, no. 7, pages 1194–1205, 2004. (Cité en page 58.)

[Borji & Itti 2013] Ali Borji et Laurent Itti. *State-of-the-art in Visual Attention Modeling.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pages 185–207, 2013. (Cité en pages 13 et 31.)

[Borji *et al.* 2011] Ali Borji, Dicky Sihite et Laurent Itti. *Computational Modeling of Top-down Visual Attention in Interactive Environments*. In British Machine Vision Conference (BMVC 2011), Dundee, UK, 2011. British Machine Vision Association. (Cité en page 18.)

[Borji *et al.* 2012] Ali Borji, Dicky N Sihite et Laurent Itti. *Quantitative Analysis of Human-Model Agreement in Visual Saliency Modeling : A Comparative Study*. IEEE Transactions on Image Processing, vol. 22, no. 1, pages 55–69, 2012. (Cité en pages 17 et 138.)

[Bovik *et al.* 1993] Alan C. Bovik, Petros Maragos et Thomas F. Quatieri. *AM-FM Energy Detection and Separation in Noise Using Multiband Energy Operators*. IEEE Transactions on Signal Processing, vol. 41, no. 12, pages 3245–3265, 1993. (Cité en page 49.)

[Bregman 1990] Albert S. Bregman. Auditory scene analysis, the perceptual organization of sound. MIT Press, Cambridge, MA, US, 1990. (Cité en pages 19, 49 et 92.)

[Buchan *et al.* 2007] Julie N Buchan, Martin Paré et Kevin G Munhall. *Spatial statistics of gaze fixations during dynamic face processing*. Social Neuroscience, vol. 2, no. 1, pages 1–13, 2007. (Cité en pages 92, 97 et 109.)

[Burr & Alais 2006] David Burr et David Alais. *Combining visual and auditory information*. Progress in Brain Research, vol. 155, pages 243–258, 2006. (Cité en page 59.)

[Buswell 1935] Guy Thomas Buswell. *How People Look at Pictures*. In A Study of the Psychology of Perception in Art. The University of Chicago Press, Chicago, USA, 1935. (Cité en pages 6, 7, 12, 13, 86, 90 et 108.)

[Calvert *et al.* 2004] G Calvert, C Spence et B E Stein. Handbook of multisensory processes. MIT Press, Cambridge, MA, USA, 2004. (Cité en page 93.)

[Carles *et al.* 1999] José Luis Carles, Isabel López Barrio et José Vicente de Lucio. *Sound influence on landscape values*. Landscape and Urban Planning, vol. 43, no. 4, pages 191–200, 1999. (Cité en page 62.)

[Carmi & Itti 2006] Ran Carmi et Laurent Itti. *Visual causes versus correlates of attentional selection in dynamic scenes*. Vision Research, vol. 46, no. 26, pages 4333–4345, 2006. (Cité en pages 11, 12, 47, 73 et 87.)

[Castelhano *et al.* 2009] Monica S. Castelhano, Michael L Mack et John M. Henderson. *Viewing task influences eye movement control during active scene perception*. Journal of Vision, vol. 9, no. 3, pages 1–15, 2009. (Cité en page 7.)

[Cerf *et al.* 2008a] Moran Cerf, Jonathan Harel, Wolfgang Einhäuser et Christof Koch. *Predicting human gaze using low-level saliency combined with face detection*. In J C Platt, D Koller, Y Singer et S Roweis, editeurs, Advances in Neural Information Processing Systems 20, pages 241–248. MIT Press, 2008. (Cité en page 113.)

[Cerf *et al.* 2008b] Moran Cerf, E Paxon Frady et Christof Koch. *Using semantic content as cues for better scanpath prediction.* In Symposium on Eye Tracking Research & Applications (ETRA), pages 143–146, Savannah, Georgia, USA, 2008. (Cité en pages 113 et 114.)

[Chamaret *et al.* 2010] C Chamaret, J C Chevet et O Le Meur. *Spatio-Temporal Combination of Saliency Maps and Eye-Tracking Assessment of Different Strategies.* In IEEE International Conference on Image Processing, pages 1077–1080, Hong Kong, 2010. (Cité en page 16.)

[Chen & Gopalakrishnan 1998] Scott Shaobing Chen et Ponani S Gopalakrishnan. *Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion.* In DARPA Broadcast News Transcription and Understanding Workshop, Virginia, USA, 1998. (Cité en page 123.)

[Chen & Yeh 2009] Yi-Chuan Chen et Su-Ling Yeh. *Catch the moment : multi-sensory enhancement of rapid visual events by sound.* Experimental Brain Research, vol. 198, pages 209–219, 2009. (Cité en page 28.)

[Chen *et al.* 2003] Li-Qun Chen, Xing Xie, Xin Fan, Wei-Ying Ma, Hong-Jiang Zhang et He-Qin Zhou. *A visual attention model for adapting images on small displays.* Multimedia Systems, vol. 9, no. 4, pages 353–364, 2003. (Cité en page 113.)

[Cheng *et al.* 2010] Shih-Sian Cheng, Hsin-Min Wang et Hsin-Chia Fu. *BIC-based Speaker Segmentation Using Divide-and-Conquer Strategies with Application to Speaker Diarization.* IEEE Transactions on Audio, Speech and Language Processing, vol. 18, no. 1, pages 141–157, 2010. (Cité en page 123.)

[Cherry 1953] E. Colin Cherry. *Some Experiments on the Recognition of Speech, with One and with Two Ears.* The Journal of the Acoustical Society of America, vol. 25, no. 5, pages 975–979, 1953. (Cité en page 20.)

[Coath *et al.* 2007] Martin Coath, Susan L Denham, Leigh Smith, Henkjan Honing, Amaury Hazan, Piotr Holonowicz et Hendrik Purwins. *An auditory model for the detection of perceptual onsets and beat tracking in singing.* In Neural Information Processing Systems, Workshop on Music Processing in the Brain, Vancouver, Canada, 2007. (Cité en pages 21 et 23.)

[Cohen 2005] Annabel J. Cohen. *How music influences the Interpretation of Film and Video : Approaches from Experimental Psychology.* In Roger A Kendall et Roger W H Savage, editeurs, Selected Reports in Ethnomusicology : Perspectives in Systematic Musicology, pages 15–36. Department of Ethnomusicology, University of California, Los Angeles, CA, USA, 2005. (Cité en page 58.)

[Cohen 2014] Annabel J. Cohen. *Film Music from the Perspective of Cognitive Science.* In David Neumeyer, editeur, The Oxford Handbook of Film Music Studies, pages 96–130. Oxford University Press, New York, NY, USA, 2014. (Cité en page 58.)

[Corneil & Munoz 1996] Brian D. Corneil et Douglas P. Munoz. *The influence of auditory and visual distractors on human orienting gaze shifts.* The Journal of neuroscience, vol. 16, no. 24, pages 8193–8207, 1996. (Cité en pages 27 et 48.)

[Corneil *et al.* 2002] Brian D. Corneil, M Van Wanrooij, D P Munoz et A. J. Van Opstal. *Auditory-visual interactions subserving goal-directed saccades in a complex scene.* Journal of Neurophysiology, vol. 88, pages 438–454, 2002. (Cité en pages 27 et 48.)

[Couronné *et al.* 2010] Thomas Couronné, Anne Guérin-Dugué, Michel Dubois, Pauline Faye et Christian Marendaz. *A statistical mixture method to reveal bottom-up and top-down factors guiding the eye-movements.* Journal of Eye Movement Research, vol. 3, no. 2, pages 1–13, 2010. (Cité en page 75.)

[Coutrot & Guyader 2013a] Antoine Coutrot et Nathalie Guyader. *Exploration of dynamic natural scenes : influence of unrelated soundtracks on eye movements.* In Kenneth Holmqvist, Fiona Mulvey et Roger Johansson, editeurs, 17th European Conference on Eye Movements, Lund, Sweden, 2013. (Cité en page 134.)

[Coutrot & Guyader 2013b] Antoine Coutrot et Nathalie Guyader. *Toward the Introduction of Auditory Information in Dynamic Visual Attention Models.* In 14th International Workshop on Image and Audio Analysis for Multimedia Interactive Services (WIAMIS 2013), Paris, France, 2013. (Cité en page 134.)

[Coutrot & Guyader 2014a] Antoine Coutrot et Nathalie Guyader. *An Audiovisual Attention Model for Natural Conversation Scenes.* In IEEE International Conference on Image Processing (ICIP), Paris, France, 2014. (Cité en pages 127 et 134.)

[Coutrot & Guyader 2014b] Antoine Coutrot et Nathalie Guyader. *How saliency, faces, and sound influence gaze in dynamic social scenes.* Journal of Vision, vol. 14, no. 8, pages 1–17, 2014. (Cité en page 134.)

[Coutrot *et al.* 2012] Antoine Coutrot, Nathalie Guyader, Gelu Ionescu et Alice Caplier. *Influence of soundtrack on eye movements during video exploration.* Journal of Eye Movement Research, vol. 5, no. 4, pages 1–10, 2012. (Cité en page 133.)

[Coutrot *et al.* 2014] Antoine Coutrot, Nathalie Guyader, Gelu Ionescu et Alice Caplier. *Video viewing : do auditory salient events capture visual attention ?* Annals of Telecommunications, vol. 69, no. 1, pages 89–97, 2014. (Cité en page 133.)

[Crouzet *et al.* 2010] Sébastien M Crouzet, Holle Kirchner et Simon J Thorpe. *Fast saccades toward faces : Face detection in just 100 ms.* Journal of Vision, vol. 10, no. 4, pages 1–17, 2010. (Cité en page 91.)

[Cutting *et al.* 2011] James E. Cutting, Jordan E. DeLong et Kaitlin L. Brunick. *Visual Activity in Hollywood Film : 1935 to 2005 and Beyond.* Psychology

of Aesthetics, Creativity, and the Arts, vol. 5, no. 2, pages 115–125, 2011. (Cité en page 11.)

[Daugman 1980] J. G. Daugman. *Two-dimensional spectral analysis of cortical receptive field profiles.* Vision Research, vol. 20, pages 847–856, 1980. (Cité en page 50.)

[Davis & Mermelstein 1980] Steven Davis et Paul Mermelstein. *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences.* IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 28, no. 4, pages 357–366, 1980. (Cité en page 123.)

[De Coensel & Botteldooren 2010] Bert De Coensel et Dick Botteldooren. *A model of saliency-based auditory attention to environmental sound.* In 20th International Congress on Acoustics, ICA 2010, Sydney, Australia, 2010. (Cité en pages 21 et 23.)

[Deleforge & Horaud 2012] Antoine Deleforge et Radu Horaud. *The Cocktail Party Robot : Sound Source Separation and Localisation with an Active Binaural Head.* In ACM/IEEE International Conference on Human Robot Interaction, pages 431–438, Boston, MA, USA, 2012. (Cité en page 21.)

[DeLong *et al.* 2012] Jordan E. DeLong, Kaitlin L. Brunick et James E. Cutting. *Film through the Human Visual System : Finding Patterns and Limits.* In J. C. Kaufman et D.K. Simonton, editeurs, Social Science of Cinema, pages 1–13. New York : Oxford University Press, 2012. (Cité en page 11.)

[Dempster *et al.* 1977] A P Dempster, N M Laird et D B Rubin. *Maximum Likelihood from Incomplete Data via the EM Algorithm.* Journal of the Royal Statistical Society. Series B (Methodological), vol. 39, no. 1, pages 1–38, 1977. (Cité en page 75.)

[Dorr *et al.* 2010] Michael Dorr, Thomas Martinetz, Karl R. Gegenfurtner et Erhardt Barth. *Variability of eye movements when viewing dynamic natural scenes.* Journal of Vision, vol. 10, no. 28, pages 1–17, 2010. (Cité en pages 9, 10, 12, 13, 46, 84 et 85.)

[Dowling *et al.* 1995] W Jay Dowling, Seyeul Kwak et Melinda W Andrews. *The time course of recognition of novel melodies.* Perception & Psychophysics, vol. 57, no. 2, pages 136–149, 1995. (Cité en page 58.)

[Doyle & Snowden 2001] Melanie C Doyle et Robert J Snowden. *Identification of visual stimuli is improved by accompanying auditory stimuli : the role of eye movements and sound.* Perception, vol. 30, pages 795–810, 2001. (Cité en page 28.)

[Driver & Spence 1998] Jon Driver et Charles Spence. *Attention and the crossmodal construction of space.* Trends in Cognitive Sciences, vol. 2, no. 7, pages 254–262, 1998. (Cité en page 94.)

[Duangudom & Anderson 2007] Varinthira Duangudom et David V. Anderson. *Using Auditory Saliency to Understand Complex Auditory Scenes.* In 15th

European Signal Processing Conference (EUSIPCO 2007), pages 1206–1210, Poznan, Poland, 2007. (Cité en page 21.)

[Duangudom Delmotte 2012] Varinthira Duangudom Delmotte. *Computational Auditory Saliency*. PhD thesis, Georgia Institute of Technology, Atlanta, GA, USA, 2012. (Cité en pages 23 et 56.)

[Dufour *et al.* 2008] André Dufour, Pascale Touzalin, Michèle Moessinger, Renaud Brochard et Olivier Després. *Visual motion disambiguation by a subliminal sound*. Consciousness and Cognition, vol. 3, no. 17, pages 790–797, 2008. (Cité en page 29.)

[Eisenbarth & Alpers 2011] H. Eisenbarth et G. W. Alpers. *Happy mouth and sad eyes : Scanning emotional facial expressions*. Emotion, vol. 11, no. 4, pages 860–865, 2011. (Cité en page 92.)

[Eisenstein 1943] Sergueï Eisenstein. The film sense. Faber and Faber, London, UK, 1943. (Cité en page 30.)

[Engelken & Stevens 1989] EJ Engelken et KW Stevens. *Saccadic eye movements in response to visual, auditory and bisensory stimuli*. Avia Space Environ Med, vol. 60, pages 762–768, 1989. (Cité en page 27.)

[Enns & Rensink 1991] J. T. Enns et R. A. Rensink. *Preattentive recovery of three-dimensional orientation from line drawings*. Psychological Review, vol. 98, no. 3, pages 335–351, 1991. (Cité en page 91.)

[Ernst & Banks 2002] Marc O. Ernst et Martin S. Banks. *Humans integrate visual and haptic information in a statistically optimal fashion*. Nature, vol. 415, pages 429–433, 2002. (Cité en pages 59 et 137.)

[Ernst & Bülthoff 2004] Marc O. Ernst et Heinrich H. Bülthoff. *Merging the senses into a robust percept*. Trends in Cognitive Sciences, vol. 8, no. 4, pages 162–169, 2004. (Cité en page 59.)

[Evangelopoulos & Maragos 2006] Georgios Evangelopoulos et Petros Maragos. *Multiband Modulation Energy Tracking for Noisy Speech Detection*. IEEE Transaction on Audio, Speech and Language Processing, vol. 14, no. 6, pages 2024–2038, 2006. (Cité en pages 21, 49 et 57.)

[Evangelopoulos *et al.* 2013] Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas et Yannis Avrithis. *Multimodal Saliency and Fusion for Movie Summarization based on Aural, Visual, and Textual Attention*. IEEE Transactions on Multimedia, vol. 15, no. 7, pages 1553–1568, 2013. (Cité en pages 32 et 49.)

[Evans & Treisman 2010] Karla K Evans et Anne M. Treisman. *Natural cross-modal mappings between visual and auditory features*. Journal of Vision, vol. 10, no. 1, pages 1–12, 2010. (Cité en page 137.)

[Follet *et al.* 2011] Brice Follet, Olivier Le Meur et Thierry Baccino. *New insights into ambient and focal visual fixations using an automatic classification algorithm*. i-Perception, vol. 2, pages 592–610, 2011. (Cité en pages 12 et 86.)

[Fong *et al.* 2003] Terrence Fong, Illah Nourbakhsh et Kerstin Dautenhahn. *A survey of socially interactive robots.* Robotics and autonomous systems, vol. 42, no. 3, pages 143–166, 2003. (Cité en page 32.)

[Foulsham & Sanderson 2013] Tom Foulsham et Lucy Anne Sanderson. *Look who's talking ? Sound changes gaze behaviour in a dynamic social scene.* Visual Cognition, vol. 21, no. 7, pages 922–944, 2013. (Cité en pages 98, 109, 111 et 120.)

[Foulsham & Underwood 2008] Tom Foulsham et Geoffrey Underwood. *What can saliency models predict about eye movements ? Spatial and sequential aspects of fixations during encoding and recognition.* Journal of Vision, vol. 8, no. 2, pages 1–17, 2008. (Cité en page 8.)

[Foulsham *et al.* 2010] Tom Foulsham, Joey T Cheng, Jessica L Tracy, Joseph Henrich et Alan Kingstone. *Gaze allocation in a dynamic situation : Effects of social status and speaking.* Cognition, vol. 117, no. 3, pages 319–331, 2010. (Cité en pages 92, 98 et 108.)

[Frank *et al.* 2009] Michael C Frank, Edward Vul et Scott P Johnson. *Development of infants' attention to faces during the first year.* Cognition, vol. 110, pages 160–170, 2009. (Cité en page 108.)

[Freeman & Driver 2008] Elliot Freeman et Jon Driver. *Direction of Visual Apparent Motion Driven Solely by Timing of a Static Sound.* Current Biology, vol. 18, pages 1262–1266, 2008. (Cité en page 29.)

[Frens *et al.* 1995] MA Frens, AJ Van Opstal et RW Van der Willigen. *Spatial and temporal factors determine audio-visual interactions in human saccadic eye movements.* Perception & Psychophysics, vol. 57, pages 802–816, 1995. (Cité en page 27.)

[Friedland *et al.* 2009] Gerald Friedland, Hayley Hung et Chuohao Yeo. *Multi-Modal Speaker Diarization of Real-World Meetings Using Compressed-Domain Video Features.* In IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP), pages 4069–4072, Taipei, Taiwan, 2009. (Cité en page 126.)

[Frisson *et al.* 2014] Christian Frisson, Nicolas Riche, Antoine Coutrot, Delestage Charles-Alexandre, Stéphane Dupont, Onur Ferhat, Nathalie Guyader, Sidi Mahmoudi Ahmed, Matei Mancas, Parag K. Mital, Alicia Prieto, François Rocca, Alexis Rochette et Willy Yvart. *Auracle : how are salient cues situated in audiovisual content ? (in press).* In International Summer Workshop on Multimodal User Interfaces (eNTERFACE), Bilbao, Spain, 2014. (Cité en page 136.)

[Fritz *et al.* 2007] Jonathan B. Fritz, Mounya Elhilali et Stephen V. David. *Auditory attention — focusing the searchlight on sound.* Current Opinion in Neurobiology, vol. 17, pages 1–19, 2007. (Cité en page 49.)

[Gabrielsson 2001] Alf Gabrielsson. *Emotions in Strong Experiences with Music.* In Patrik N. Juslin et John A. Sloboda, editeurs, Music and Emotion : Theory

and Research. Series in affective science., pages 431–449. Oxford University Press, New York, NY, USA, 2001. (Cité en page 58.)

[Gauss 1809] Johann Carl Friedrich Gauss. Théorie du mouvement des corps célestes parcourant des sections coniques autour du soleil. Perthes, F and Besse, I. H., Hambourg, Allemagne, 1809. (Cité en page 76.)

[Gautier & Le Meur 2012] Josselin Gautier et Olivier Le Meur. *A Time-Dependent Saliency Model Combining Center and Depth Biases for 2D and 3D Viewing Conditions*. Cognitive Computation, vol. 4, pages 1–16, 2012. (Cité en pages 13, 75 et 79.)

[Gebhard & Mowbray 1959] JW Gebhard et GH Mowbray. *On discriminating the rate of visual flicker and auditory flutter*. American Journal of Psychology, vol. 72, pages 521–528, 1959. (Cité en page 29.)

[Gebrekidan Gebre *et al.* 2013] Binyam Gebrekidan Gebre, Peter Wittenburg et Tom Heskes. *The Gesturer Is the Speaker*. In IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP), Vancouver, Canada, 2013. (Cité en pages 122 et 126.)

[Giannakopoulos 2010] Theodoros Giannakopoulos. *Silence removal in speech signals*. http ://www.mathworks.com, 2010. (Cité en page 122.)

[Goferman *et al.* 2012] Stas Goferman, Lihi Zelnik-Manor et Ayellet Tal. *Context-Aware Saliency Detection*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 10, pages 1915–1926, 2012. (Cité en page 113.)

[Goldring *et al.* 1996] JE Goldring, MC Dorris, BD Corneil, PA Ballantyne et DP Munoz. *Combined eye-head gaze shifts to visual and auditory targets in humans*. Experimental Brain Research, vol. 111, pages 68–78, 1996. (Cité en page 27.)

[Goldstein *et al.* 2007] Robert B. Goldstein, Russell L. Woods et Eli Peli. *Where people look when watching movies : Do all viewers look at the same place ?* Computers in Biology and Medicine, vol. 37, no. 7, pages 957–964, 2007. (Cité en pages 12 et 46.)

[Gosselin & Schyns 2001] Frédéric Gosselin et Philippe G Schyns. *Bubbles : a technique to reveal the use of information in recognition tasks*. Vision Research, vol. 41, pages 2261–2271, 2001. (Cité en page 92.)

[Green *et al.* 1991] K P. Green, P K. Kuhl, A N. Meltzoff et E B. Stevens. *Integrating speech information across talkers, gender, and sensory modality : female faces and male voices in the McGurk effect*. Perception & Psychophysics, vol. 50, pages 524–536, 1991. (Cité en page 94.)

[Haaf & Bell 1967] R. A. Haaf et R. Q. Bell. *A facial dimension in visual discrimination by human infants*. Child Development, vol. 38, no. 3, pages 893–899, 1967. (Cité en page 90.)

[Hadizadeh *et al.* 2012] H. Hadizadeh, M. J. Enriquez et I. V. Bajić. *Eye-tracking database for a set of standard video sequences*. IEEE Transactions on Image Processing, vol. 21, no. 2, pages 898–903, 2012. (Cité en page 8.)

[Harding & Bloj 2010] Glen Harding et Marina Bloj. *Real and predicted influence of image manipulations on eye movements during scene recognition.* Journal of Vision, vol. 10, no. 2, pages 1–17, 2010. (Cité en page 8.)

[Hasson *et al.* 2008a] Uri Hasson, Ohad Landesman, Barbara Knappmeyer, Ignacio Vallines, Nava Rubin et David J. Heeger. *Neurocinematics : The Neuroscience of Film.* Projections, vol. 2, no. 1, pages 1–26, 2008. (Cité en pages x et 8.)

[Hasson *et al.* 2008b] Uri Hasson, Eunice Yang, Ignacio Vallines, David J. Heeger et Nava Rubin. *A Hierarchy of Temporal Receptive Windows in Human Cortex.* The Journal of neuroscience, vol. 28, no. 10, pages 2539–2550, 2008. (Cité en page 46.)

[Hastie *et al.* 2009] Trevor Hastie, Robert Tibshirnani et Jerome Friedman. The Elements of Statistical Learning : Data Mining, Inference, and Prediction. Springer Series in Statistics. Springer-Verlag New York Inc, 2009. (Cité en page 77.)

[Haxby *et al.* 2000] James V Haxby, Elizabeth A Hoffman et M Ida Gobbini. *The distributed human neural system for face perception.* Trends in Cognitive Sciences, vol. 4, no. 6, pages 223–233, 2000. (Cité en page 90.)

[Henderson & Hollingworth 1998] John M. Henderson et Andrew Hollingworth. *Eye Movements During Scene Viewing : an overview.* In G Underwood, editeur, Eye Guidance in Reading and Scene Perception, pages 269–290. Elsevier Science Ltd, Oxford, UK, 1998. (Cité en page 7.)

[Hershey & Movellan 1999] John Hershey et Javier Movellan. *Audio-Vision : Using Audio-Visual Synchrony to Locate Sounds.* In Advances in Neural Information Processing Systems 12 (NIPS 1999), pages 813–819, Denver, Colorado, USA, 1999. MIT Press. (Cité en page 121.)

[Hershler & Hochstein 2005] Orit Hershler et Shaul Hochstein. *At first sight : A high-level pop out effect for faces.* Vision Research, vol. 45, pages 1707–1724, 2005. (Cité en pages 64 et 91.)

[Hidaka *et al.* 2009] Souta Hidaka, Yuko Manaka, Wataru Teramoto, Yoichi Sugita, Ryota Miyauchi, Jiro Gyoba, Yôiti Suzuki et Yukio Iwaya. *Alternation of Sound Location Induces Visual Motion Perception of a Static Object.* PLoS ONE, vol. 4, no. 12, pages 1–6, 2009. (Cité en page 29.)

[Hirvenkari *et al.* 2013] Lotta Hirvenkari, Johanna Ruusuvori, Veli-Matti Saarinen, Maari Kivioja, Anssi Peräkylä et Riitta Hari. *Influence of Turn-Taking in a Two-Person Conversation on the Gaze of a Viewer.* PLoS ONE, vol. 8, no. 8, pages 1–6, 2013. (Cité en pages 98 et 109.)

[Ho-Phuoc *et al.* 2010] Tien Ho-Phuoc, Nathalie Guyader et Anne Guerin-Dugue. *A Functional and Statistical Bottom-Up Saliency Model to Reveal the Relative Contributions of Low-Level Visual Guiding Factors.* Cognitive Computation, vol. 2, no. 4, pages 344–359, 2010. (Cité en page 75.)

[Ho-Phuoc *et al.* 2012] Tien Ho-Phuoc, Nathalie Guyader, Frédéric Landragin et Anne Guérin-Dugué. *When viewing natural scenes, do abnormal colours impact on spatial or temporal parameters of eye movements ?* Journal of Vision, vol. 12, no. 2, pages 1–13, 2012. (Cité en page 40.)

[Hoffman 1998] James E. Hoffman. *Visual Attention and Eye Movements.* In H Pashler, editeur, Attention, pages 119–154. University College London Press, London, 1998. (Cité en page 4.)

[Hotelling 1936] Harold Hotelling. *Relations between two sets of variates.* Biometrika, vol. 28, no. 3–4, pages 321–377, 1936. (Cité en page 135.)

[Hubel & Wiesel 1959] David H Hubel et Torsten N Wiesel. *Receptive Fields of Single Neurones in the Cat's Striate Cortex.* The Journal of Physiology, vol. 148, pages 574–591, 1959. (Cité en page 90.)

[Hughes *et al.* 1994] HC Hughes, PA Reuter-Lorenz, G Nozawa et R Fendrich. *Visual- auditory interactions in sensorimotor processing : saccades versus manual responses.* Journal of Experimental Psychology, vol. 20, pages 131–153, 1994. (Cité en page 27.)

[Hui-wen Hsiao & Cottrell 2008] Janet Hui-wen Hsiao et Garrison W. Cottrell. *Two fixations suffice in face recognition.* Psychological Science, vol. 19, no. 10, pages 998–1006, 2008. (Cité en page 92.)

[Hung *et al.* 2008] Hayley Hung, Daniel Gatica-Perez, Yan Huang et Gerald Friedland. *Estimating the Dominant Person in Multi-Party Conversations Using Speaker Diarization Strategies.* In IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP), pages 2197–2200, Las Vegas, Nevada, USA, 2008. (Cité en page 126.)

[Ionescu *et al.* 2009] Gelu Ionescu, Nathalie Guyader et Anne Guérin-Dugué. *SoftEye software*, 2009. IDDN.FR.001.200017.000.S.P.2010.003.31235. (Cité en page 37.)

[Isola *et al.* 2011] Phillip Isola, Jianxiong Xiao, Antonio Torralba et Aude Oliva. *What makes an image memorable ?* In IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007), pages 145–152, 2011. (Cité en page 85.)

[Itier *et al.* 2007] RJ Itier, C Villate et JD. Ryan. *Eyes always attract attention but gaze orienting is task-dependent : evidence from eye movement monitoring.* Neuropsychologia, vol. 45, no. 5, pages 1019–1028, 2007. (Cité en page 92.)

[Itti & Koch 2000] Laurent Itti et Christof Koch. *A saliency-based search mechanism for overt and covert shifts of visual attention.* Vision Research, vol. 40, pages 1489–1506, 2000. (Cité en pages 87 et 92.)

[Itti *et al.* 1998] Laurent Itti, Christof Koch et Ernst Niebur. *A model of saliency-based visual attention for rapid scene analysis.* IEEE Transactions on Pattern Analysis and Machine Intellingence, vol. 20, no. 11, pages 1254–1259, 1998. (Cité en pages 14 et 113.)

[Itti 2005] Laurent Itti. *Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes.* Visual Cognition, vol. 12, no. 6, pages 1093–1123, 2005. (Cité en pages 14 et 73.)

[Jansen *et al.* 2009] Lina Jansen, Selim Onat et Peter König. *Influence of disparity on fixation and saccades in free viewing of natural scenes.* Journal of Vision, vol. 9, no. 1, pages 1–19, 2009. (Cité en pages 84, 85 et 87.)

[Judd *et al.* 2011] Tilke Judd, Frédo Durand et Antonio Torralba. *Fixations on Low-Resolution Images.* Journal of Vision, vol. 11, no. 4, pages 1–23, 2011. (Cité en pages 84 et 85.)

[Judd *et al.* 2012] Tilke Judd, Frédo Durand et Antonio Torralba. *A Benchmark of Computational Models of Saliency to Predict Human Fixations.* Rapport technique MIT-CSAIL-TR-2012-001, MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA, 2012. (Cité en page 138.)

[Juslin & Laukka 2004] Patrik N. Juslin et Petri Laukka. *Expression, Perception, and Induction of Musical Emotions : A Review and a Questionnaire Study of Everyday Listening.* Journal of New Music Research, vol. 33, no. 3, pages 217–238, 2004. (Cité en page 58.)

[Kadunce *et al.* 1997] D.C. Kadunce, J.W. Vaughan, M.T. Wallace, G. Benedek et B.E. Stein. *Mechanisms of within-and cross-modality suppression in the superior colliculus.* Journal of Neurophysiology, vol. 78, no. 6, pages 2834–2847, 1997. (Cité en page 25.)

[Kaiser 1990] Jim Kaiser. *On a simple algorithm to calculate the "energy" of a signal.* In International Conference on Acoustics, Speech, and Signal Processing, ICASSP-90, volume 1, pages 381–384, Albuquerque, NM, USA, 1990. (Cité en page 50.)

[Kalinli & Narayanan 2007] Ozlem Kalinli et Shrikanth Narayanan. *A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech.* In Eighth Annual Conference of the International Speech Communication Association, pages 1941–1944, Antwerp, Belgium, 2007. (Cité en pages 21 et 23.)

[Kanwisher *et al.* 1997] Nancy Kanwisher, Josh McDermott et Marvin M Chun. *The Fusiform Face Area : A Module in Human Extrastriate Cortex Specialized for Face Perception.* The Journal of neuroscience, vol. 17, no. 11, pages 4302–4311, 1997. (Cité en page 90.)

[Kavak *et al.* 2013] Yasin Kavak, Erkut Erdem et Aykut Erdem. *Visual saliency estimation by integrating features using multiple kernel learning.* In IEEE Conference on Computer Vision and Pattern Recognition, Portland, Oregon, USA, 2013. (Cité en page 16.)

[Kaya & Elhilali 2012] Emine Merve Kaya et Mounya Elhilali. *A temporal saliency map for modeling auditory attention.* In 46th Annual Conference on Information Sciences and Systems (CISS), Princeton, NJ, USA, 2012. (Cité en pages 21 et 23.)

[Kayser *et al.* 2005] Christoph Kayser, Christopher I. Petkov, Michael Lippert et Nikos K. Logothetis. *Mechanisms for allocating auditory attention : an auditory saliency map.* Current Biology, vol. 15, pages 1943–1947, 2005. (Cité en pages 21, 22 et 49.)

[Kidron *et al.* 2007] Einat Kidron, Yoav Y. Schechner et Michael Elad. *Cross-Modal Localization via Sparsity.* IEEE Transaction on Signal Processing, vol. 55, no. 4, pages 1390–1404, 2007. (Cité en page 136.)

[Klein 1980] R. M. Klein. *Does oculomotor readiness mediate cognitive control of visual attention ?* In R. S. Nickerson, editeur, Attention and performance VIII, pages 259–276. Lawrence Erlbaum, Hillsdale, NJ, 1980. (Cité en page 5.)

[Krieger *et al.* 2000] Gerhard Krieger, Ingo Rentschler, Gert Hauske, Kerstin Schill et Christoph Zetzsche. *Object and scene analysis by saccadic eye-movements : an investigation with higher-order statistics.* Spatial Vision, vol. 13, no. 2,3, pages 201–214, 2000. (Cité en page 73.)

[Kullback & Leibler 1951] S. Kullback et R. A. Leibler. *On Information and Sufficiency.* The Annals of Mathematical Statistics, vol. 22, no. 1, pages 79–86, 1951. (Cité en page 17.)

[Lansing & McConkie 2003] Charissa R Lansing et George W McConkie. *Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences.* Perception & Psychophysics, vol. 65, no. 4, pages 536–552, 2003. (Cité en page 97.)

[Lavie 2005] Nilli Lavie. *Distracted and confused ? : Selective attention under load.* Trends in Cognitive Sciences, vol. 9, no. 2, pages 75–82, 2005. (Cité en page 97.)

[Le Meur & Baccino 2013] Olivier Le Meur et Thierry Baccino. *Methods for comparing scanpaths and saliency maps : strengths and weaknesses.* Behavior Research Methods, vol. 45, no. 1, pages 251–266, 2013. (Cité en pages 17, 40, 74 et 104.)

[Le Meur *et al.* 2007] Olivier Le Meur, Patrick Le Callet et Dominique Barba. *Predicting visual fixations on video based on low-level visual features.* Vision Research, vol. 47, pages 2483–2498, 2007. (Cité en page 40.)

[Legendre 1805] Adrien-Marie Legendre. *Appendice sur la méthodes des moindres quarrés.* In Nouvelles méthodes pour la détermination des orbites des comètes, pages 72–80. Firmin-Didot, Paris, France, 1805. (Cité en page 76.)

[Lettvin *et al.* 1959] J. Y. Lettvin, H. R. Maturana, W. S. McCulloch et W. H. Pitts. *What the Frog's Eye Tells the Frog's Brain.* Proceedings of the Institute of Radio Engineers, vol. 47, no. 11, pages 1940–1951, 1959. (Cité en page 5.)

[Levenshtein 1966] V. I. Levenshtein. *Binary codes capable of correcting deletions, insertions, and reversals.* Journal of Vision, vol. 10, no. 8, pages 707–710, 1966. (Cité en page 103.)

[Li *et al.* 2010] Jia Li, Yonghong Tian, Tiejun Huang et Wen Gao. *Probabilistic Multi-Task Learning for Visual Saliency Estimation in Video.* International Journal of Computer Vision, vol. 90, no. 2, pages 150–165, 2010. (Cité en page 59.)

[Ma *et al.* 2005] Yu-Fei Ma, Xian-Sheng Hua, Lie Lu et Hong-Jiang Zhang. *A Generic Framework of User Attention Model and Its Application in Video Summarization.* IEEE Transactions on Multimedia, vol. 7, no. 5, pages 907–919, 2005. (Cité en page 113.)

[MacDonald & Tatler 2013] J MacDonald et Benjamin W. Tatler. *Do as eye say : Gaze cueing and language in a real-world social interaction.* Journal of Vision, vol. 13, no. 4, pages 1–12, 2013. (Cité en page 98.)

[Mancas & Le Meur 2013] Matei Mancas et Olivier Le Meur. *Memorability of Natural Scenes : the Role of Attention.* In IEEE International Conference on Image Processing (ICIP), Melbourne, Australia, 2013. (Cité en pages 84 et 85.)

[Mannan *et al.* 1995] S Mannan, KH Ruddock et DS Wooding. *Automatic control of saccadic eye movements made in visual inspection of briefly presented 2-D images.* Spatial vision, vol. 9, no. 3, pages 363–386, 1995. (Cité en pages 10 et 85.)

[Mannan *et al.* 1996] Sabira K Mannan, Keith H Ruddock et David S Wooding. *The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images.* Spatial vision, vol. 10, no. 3, pages 165–188, 1996. (Cité en page 11.)

[Marat *et al.* 2009] Sophie Marat, Tien Ho-Phuoc, Lionel Granjon, Nathalie Guyader, Denis Pellerin et Anne Guérin-Dugué. *Modelling Spatio-Temporal Saliency to Predict Gaze Direction for Short Videos.* International Journal of Computer Vision, vol. 82, no. 3, pages 231–243, 2009. (Cité en pages 14, 15, 79, 87, 125 et 129.)

[Marat *et al.* 2013] Sophie Marat, Anis Rahman, Denis Pellerin, Nathalie Guyader et Dominique Houzet. *Improving Visual Saliency by Adding 'Face Feature Map' and 'Center Bias'.* Cognitive Computation, vol. 5, no. 1, pages 63–75, 2013. (Cité en pages 13, 15, 16, 113 et 114.)

[Marendaz *et al.* 2007] Christian Marendaz, Nathalie Guyader et Jennifer Malsert. *« Ce que l'oeil nous dit du cerveau ». Fonctions exécutives, saccades oculaires & Neuropsychologie – Neuropsychiatrie.* Revue de Neuropsychologie, vol. 17, no. 1, pages 1–35, 2007. (Cité en page 9.)

[McCowan *et al.* 2005] I McCowan, J Carletta, W Kraaij, W Ashby, S Bourban, M Flynn, M Guillemot, T Hain, J Kaldec, V Karaiskos, M Kronenthal, G Lathoud, M Lincoln, A Lisowska, W Post, D Reidsma et P Wellner. *The AMI Meeting Corpus.* In International Conference on Methods and Techniques in Behavioral Research, Wageningen, The Netherlands, 2005. (Cité en page 115.)

[McGurk & MacDonald 1976] H McGurk et J MacDonald. *Hearing lips and seeing voices.* Nature, vol. 264, pages 746–748, 1976. (Cité en page 94.)

[McNeill 1985] David McNeill. *So you think gestures are nonverbal ?* Psychological Review, vol. 92, no. 3, pages 350–371, 1985. (Cité en page 122.)

[Mehoudar *et al.* 2014] Eyal Mehoudar, Joseph Arizpe, Chris I Baker et Galit Yovel. *Faces in the eye of the beholder : Unique and stable eye scanning patterns of individual observers.* Journal of Vision, vol. 14, no. 7, pages 1–11, 2014. (Cité en page 92.)

[Meredith & Stein 1986] M.A. Meredith et B.E. Stein. *Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration.* Journal of Neurophysiology, vol. 56, no. 3, pages 640–662, 1986. (Cité en pages 25 et 57.)

[Meredith *et al.* 1987] M. Alex Meredith, James W. Nemitz et Barry E. Stein. *Determinants of Multisensory Integration in Superior Colliculus Neurons. I. Temporal Factors.* The Journal of Neuroscience, vol. 7, no. 10, pages 3215–3229, 1987. (Cité en page 57.)

[Meyer & Wuerger 2001] G. F. Meyer et S. M. Wuerger. *Cross-modal integration of auditory and visual motion signals.* NeuroReport, vol. 12, no. 11, pages 2557–2560, 2001. (Cité en page 29.)

[Milanese *et al.* 1994] R Milanese, H Wechsler et S Gill. *Integration of bottom-up and top-down cues for visual attention using non-linear relaxation.* Computer Vision and Pattern Recognition, 1994. (Cité en page 18.)

[Miller 1982] JO Miller. *Divided attention : evidence for coactivation with redundant signals.* Cognitive psychology, vol. 14, pages 247–279, 1982. (Cité en page 27.)

[Mills *et al.* 2011] Mark Mills, Hollingworth, Andrew, Stefan Van der Stigchel, Lesa Hoffman et Michael D Dodd. *Examining the influence of task set on eye movements and fixations.* Journal of Vision, vol. 11, no. 8, pages 1–15, 2011. (Cité en pages 7, 12 et 86.)

[Mirghafori & Wooters 2006] Nikki Mirghafori et Chuck Wooters. *Nuts and Flakes : a Study of Data Characteristics in Speaker Diarization.* In IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP), pages 1017–1020, Toulouse, France, 2006. (Cité en page 127.)

[Mital *et al.* 2010] Parag K. Mital, Tim J. Smith, Robin L. Hill et John M. Henderson. *Clustering of Gaze During Dynamic Scene Viewing is Predicted by Motion.* Cognitive Computation, vol. 3, no. 1, pages 5–24, 2010. (Cité en pages 12, 46, 47, 64, 73, 84, 85 et 108.)

[Muddamsetty *et al.* 2013] Satya M Muddamsetty, Désiré Sidibé, Alain Trémeau et Fabrice Mériaudeau. *A Performance Evaluation of Fusion Techniques for Spatio-Temporal Saliency Detection in Dynamic Scenes.* In IEEE International Conference on Image Processing, Melbourne, Australia, 2013. (Cité en page 16.)

[Munhall *et al.* 1996] K G. Munhall, P. Gribble, M. Sacco et M. Ward. *Temporal constraints on the McGurk effect.* Perception & Psychophysics, vol. 58, pages 351–362, 1996. (Cité en page 94.)

[Nahorna *et al.* 2012] Olha Nahorna, Frédéric Berthommier et Jean-Luc Schwartz. *Binding and unbinding the auditory and visual streams in the McGurk effect.* The Journal of the Acoustical Society of America, vol. 132, no. 2, pages 1061–1077, 2012. (Cité en page 110.)

[Navarra *et al.* 2010] Jordi Navarra, Agnès Alsius, Salvador Soto-Faraco et Charles Spence. *Assessing the role of attention in the audiovisual integration of speech.* Information Fusion, vol. 11, no. 1, pages 4–11, 2010. (Cité en page 94.)

[Ng 2004] Andrew Y Ng. *Feature selection, L1 vs. L2 regularization, and rotational invariance.* In International Conference on Machine Learning (ICML '04), Banff, Alberta, Canada, 2004. (Cité en page 77.)

[Noulas *et al.* 2012] Athanasios K. Noulas, Gwenn Englebienne et Ben J. A. Kröse. *Multimodal Speaker Diarization.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 1, pages 79–93, 2012. (Cité en page 122.)

[Noulas 2010] Athanasios K Noulas. *Audiovisual fusion for speaker diarization.* PhD thesis, University of Amsterdam, Amsterdam, the Netherlands, 2010. (Cité en page 123.)

[Nozawa *et al.* 1994] G Nozawa, PA Reuter-Lorenz et HC Hughes. *Parallel and serial processes in the human oculomotor system : bimodal integration and express saccades.* Biological Cybernetics, vol. 72, pages 19–34, 1994. (Cité en page 27.)

[Nyström & Holmqvist 2008] Marcus Nyström et Kenneth Holmqvist. *Semantic Override of Low-level Features in Image Viewing – Both Initially and Overall.* Journal of Eye Movement Research, vol. 2, no. 2, pages 1–11, 2008. (Cité en page 113.)

[Olivers & Van der Burg 2008] Christian N L Olivers et Erik Van der Burg. *Bleeping you out of the blink : sound saves vision from oblivion.* Brain Research, vol. 1242, pages 191–199, 2008. (Cité en page 28.)

[Onat *et al.* 2007] Selim Onat, Klaus Libertus et Peter König. *Integrating audiovisual information for the control of overt attention.* Journal of Vision, vol. 7, no. 10, pages 1–16, 2007. (Cité en pages 29, 30, 48 et 62.)

[Otero-Millan *et al.* 2011] Jorge Otero-Millan, Stephen L Macknik, Apollo Robbins et Susana Martinez-Conde. *Stronger misdirection in curved than in straight motion.* Frontiers in Human Neuroscience, vol. 5, pages 1–4, 2011. (Cité en page 3.)

[Pannasch *et al.* 2008] Sebastian Pannasch, Jens R. Helmert, Ann-Katrin Herbold, Katharina Roth et Walter Henrik. *Visual Fixation Durations and Saccade Amplitudes : Shifting Relationship in a Variety of Conditions.* Journal of Eye Movement Research, vol. 2, no. 4, pages 1–19, 2008. (Cité en pages 12 et 86.)

[Paré *et al.* 2003] Martin Paré, Rebecca C. Richler, Martin ten Hove et K. G. Munhall. *Gaze behavior in audiovisual speech perception : The influence of ocular fixations on the McGurk effect.* Perception & Psychophysics, vol. 65, no. 4, pages 553–567, 2003. (Cité en page 97.)

[Parkhurst & Niebur 2003] Derrick J Parkhurst et Ernst Niebur. *Scene content selected by active vision.* Spatial Vision, vol. 16, no. 2, pages 125–154, 2003. (Cité en page 73.)

[Parkhurst *et al.* 2002] Derrick Parkhurst, Klinton Law et Ernst Niebur. *Modeling the role of salience in the allocation of overt visual attention.* Vision Research, vol. 42, pages 107–123, Janvier 2002. (Cité en pages 13 et 87.)

[Perrott *et al.* 1990] DR Perrott, K Saberi, K Brown et TZ Strybel. *Auditory psychomotor coordination and visual search performance.* Perception & Psychophysics, vol. 48, pages 214–226, 1990. (Cité en page 27.)

[Peters *et al.* 2005] Robert J. Peters, Asha Iyer, Laurent Itti et Christof Koch. *Components of bottom-up gaze allocation in natural images.* Vision Research, vol. 45, pages 2397–2416, 2005. (Cité en pages 17 et 73.)

[Potter 1976] Mary C Potter. *Short-term conceptual memory for pictures.* Journal of Experimental Psychology : Human Learning and Memory, vol. 2, no. 5, pages 509–522, 1976. (Cité en page 83.)

[Queste-Devillez 2014] Hélène Queste-Devillez. *Etude des processus attentionnels mis en jeu lors de l'exploration de scènes naturelles : enregistrement conjoint des mouvements oculaires et de l'activité EEG.* PhD thesis, Université de Grenoble-Alpes, Grenoble, France, 2014. (Cité en page 75.)

[Quigley *et al.* 2008] Cliodhna Quigley, Selim Onat, Sue Harding, Martin Cooke et Peter König. *Audio-visual integration during overt visual attention.* Journal of Eye Movement Research, vol. 1, no. 2, pages 1–17, 2008. (Cité en pages 30, 40, 48 et 62.)

[Raab 1962] D Raab. *Statistical facilitation of simple reaction times.* Transactions of the New York Academy of Sciences, vol. 24, pages 574–590, 1962. (Cité en page 27.)

[Rahman *et al.* 2014] Anis Rahman, Denis Pellerin et Dominique Houzet. *Influence of number, location and size of faces on gaze in video.* Journal of Eye Movement Research, vol. 7, no. 2, pages 1–11, 2014. (Cité en page 114.)

[Rapantzikos & Evangelopoulos 2007] K Rapantzikos et G Evangelopoulos. *An audio-visual saliency model for movie summarization.* Institute of Electrical and Electronics Engineers, page 425, 2007. (Cité en page 32.)

[Rayner 1998] Keith Rayner. *Eye Movements In Reading And Information Processing : 20 years of research.* Psychological Bulletin, vol. 124, no. 3, pages 372–422, 1998. (Cité en page 6.)

[Recanzone 2003] GH Recanzone. *Auditory influences on visual temporal rate perception.* Journal of Neurophysiology, vol. 89, pages 1078–1093, 2003. (Cité en page 29.)

[Recanzone 2009] Gregg H Recanzone. *Interactions of auditory and visual stimuli in space and time.* Hearing Research, vol. 258, no. 1-2, pages 89–99, 2009. (Cité en page 57.)

[Richardson & Dale 2005] Daniel C Richardson et Rick Dale. *Looking To Understand : The Coupling Between Speakers' and Listeners' Eye Movements and its Relationship to Discourse Comprehension.* Cognitive Science, vol. 29, pages 39–54, 2005. (Cité en page 98.)

[Richardson *et al.* 2008] D Richardson, R Dale et K Shockley. *Synchrony and swing in conversation : coordination, temporal dynamics, and communication.* In I Wachsmuth, M Lenzen et G Knoblich, editeurs, Embodied Communication, pages 75–94. Oxford University Press, New York, 2008. (Cité en page 109.)

[Richardson *et al.* 2009] Daniel C Richardson, Rick Dale et John M Tomlinson. *Conversation, Gaze Coordination, and Beliefs About Visual Context.* Cognitive Science, vol. 33, pages 1468–1482, 2009. (Cité en page 91.)

[Richardson *et al.* 2012] Daniel C Richardson, Chris N H Street, Joanne Y M Tan, Natasha Z Kirkham, Merrit A Hoover et Arezou Ghane Cavanaugh. *Joint perception : gaze and social context.* Frontiers in Human Neuroscience, vol. 6, pages 1–8, 2012. (Cité en page 98.)

[Riche *et al.* 2013a] Nicolas Riche, Matthieu Duvinage, Matei Mancas, Bernard Gosselin et Thierry Dutoit. *A study of parameters affecting visual saliency assessment.* In Proceedings of the 6th International Symposium on Attention in Cognitive Systems (ISACS'13), Beijing, China, 2013. (Cité en page 138.)

[Riche *et al.* 2013b] Nicolas Riche, Matthieu Duvinage, Matei Mancas, Bernard Gosselin et Thierry Dutoit. *Saliency and Human Fixations : State-of-the-art and Study of Comparison Metrics.* In Proceedings of the 14th International Conference on Computer Vision (ICCV 2013), Sydney, Australia, 2013. (Cité en pages 17 et 138.)

[Rizzolati *et al.* 1987] Giacomo Rizzolati, Lucia Riggio, Isabella Dascola et Carlo Umiltá. *Reorienting attention across the horizontal and vertical meridians : Evidence in favor of a premotor theory of attention.* Neuropsychologia, vol. 25, no. 1, pages 31–40, 1987. (Cité en page 4.)

[Ross *et al.* 2007] Lars A Ross, Dave Saint-Amour, Victoria M Leavitt, Daniel C Javitt et John J Foxe. *Do You See What I Am Saying ? Exploring Visual Enhancement of Speech Comprehension in Noisy Environments.* Cerebral Cortex, vol. 17, pages 1147–1153, 2007. (Cité en page 92.)

[Ruesch *et al.* 2008] Jonas Ruesch, Manuel Lopes, Alexandre Bernardino, Jonas Hörnstein, José Santos-Victor et Rolf Pfeifer. *Multimodal Saliency-Based Bottom-Up Attention, A Framework for the Humanoid Robot iCub.* In IEEE International Conference on Robotics and Automation, pages 962–967, Pasadena, CA, USA, 2008. (Cité en pages 21, 32, 33 et 48.)

[Rummukainen *et al.* 2014] Olli Rummukainen, Jenni Radun, Toni Virtanen et Ville Pulkki. *Categorization of Natural Dynamic Audiovisual Scenes.* PLoS ONE, vol. 9, no. 5, pages 1–14, 2014. (Cité en page 63.)

[Schauerte & Stiefelhagen 2013] Boris Schauerte et Rainer Stiefelhagen. *"Wow!" Bayesian Surprise for Salient Acoustic Event Detection.* In IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP), Vancouver, Canada, 2013. (Cité en page 21.)

[Schauerte *et al.* 2011] Boris Schauerte, Benjamin Kühn, Kristian Kroschel et Rainer Stiefelhagen. *Multimodal Saliency-based Attention for Object-based Scene Analysis.* In International Conference on Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ, pages 1173–1179, San Francisco, CA, USA, 2011. (Cité en page 21.)

[Schwarz 1978] Gideon E Schwarz. *Estimating the dimension of a model.* Annals of Statistics, vol. 6, no. 2, pages 461–464, 1978. (Cité en page 77.)

[Sekuler *et al.* 1997] R. Sekuler, A. B. Sekuler et R. Lau. *Sound alters visual motion perception.* Nature, vol. 385, no. 6614, page 308, 1997. (Cité en page 29.)

[Senkowski *et al.* 2008] Daniel Senkowski, Dave Saint-Amour, Thomas Gruber et John J Foxe. *Look who's talking : The deployment of visuo-spatial attention during multisensory speech processing under noisy environmental conditions.* Neuroimage, vol. 43, pages 379–387, 2008. (Cité en page 94.)

[Séré *et al.* 2000] Boubakar Séré, Christian Marendaz et Jeanny Hérault. *Nonhomogeneous resolution of images of natural scenes.* Perception, vol. 29, no. 12, pages 1403–1412, 2000. (Cité en page 5.)

[Shams & Kim 2010] Ladan Shams et Robyn Kim. *Crossmodal influences on visual perception.* Physics of Life Reviews, vol. 7, no. 3, pages 269–284, 2010. (Cité en page 137.)

[Shimamura 2013] Arthur P Shimamura. Psychocinematics : Exploring Cognition at the Movies. Oxford University Press, New York, USA, 2013. (Cité en pages x et 11.)

[Shipley 1964] T Shipley. *Auditory flutter-driving of visual flicker.* Science, vol. 145, pages 1328–1330, 1964. (Cité en page 29.)

[Sjöstrand *et al.* 2012] Karl Sjöstrand, Line Harder Clemmensen, Rasmus Larsen et Bjarne Ersbøll. *SpaSM : A Matlab Toolbox for Sparse Statistical Modeling.* Journal of Statistical Software, pages 1–24, 2012. (Cité en page 77.)

[Slaney & Covell 2000] Malcom Slaney et Michele Covell. *FaceSync : A linear operator for measuring synchronization of video facial images and audio tracks.* In Advances in Neural Information Processing Systems 13 (NIPS 2000), pages 814–820, Denver, CO, USA, 2000. (Cité en page 135.)

[Sloboda & O'Neill 2001] J. A. Sloboda et S. A O'Neill. *Emotions in everyday listening to music.* In Patrik N. Juslin et John A. Sloboda, editeurs, Music and Emotion : Theory and Research. Series in affective science., pages 413–429. Oxford University Press, New York, NY, USA, 2001. (Cité en page 58.)

[Smith & Mital 2013] Tim J. Smith et Parag K. Mital. *Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes.* Journal of Vision, vol. 13, no. 8, pages 1–24, 2013. (Cité en pages 7, 9, 12, 13, 47, 84, 85 et 87.)

[Smith *et al.* 2005] M. L. Smith, F. Gosselin et P. G. Cottrell G. W.and Schyns. *Transmitting and decoding facial expressions.* Psychological Science, vol. 16, pages 184–189, 2005. (Cité en page 92.)

[Smith *et al.* 2012] Tim J. Smith, Daniel Levin et James E. Cutting. *A Window on Reality : Perceiving Edited Moving Images.* Current Directions in Psychological Science, vol. 21, no. 2, pages 107–113, 2012. (Cité en page 46.)

[Smith 2012] Tim J. Smith. *The Attentional Theory of Cinematic Continuity.* Projections, vol. 6, no. 1, pages 1–50, 2012. (Cité en page 11.)

[Smith 2013] Tim J. Smith. *Watching You Watch Movies : Using Eye Tracking to Inform Cognitive Film Theory.* In Arthur P Shimamura, editeur, Psychocinematics : Exploring Cognition at the Movies, pages 165–191. New York : Oxford University Press, 2013. (Cité en pages 13 et 84.)

[Smith 2014] Tim J. Smith. *Audiovisual Correspondences in Sergei Eisenstein's Alexander Nevsky : A Case Study in Viewer Attention.* In Ted Nannicelli et Paul Taberham, editeurs, Cognitive Media Theory, pages 85–105. AFI Film Readers, 2014. (Cité en pages 30 et 31.)

[Song *et al.* 2013] Guanghan Song, Denis Pellerin et Lionel Granjon. *Different types of sounds influence gaze differently in videos.* Journal of Eye Movement Research, vol. 6, no. 4, pages 1–13, 2013. (Cité en pages 40 et 88.)

[Song 2013] Guanghan Song. *Effect of sound in videos on gaze : Contribution to audio-visual saliency modeling.* PhD thesis, Université de Grenoble-Alpes, 2013. (Cité en page 31.)

[Spence & Driver 1997] Charles Spence et Jon Driver. *Audiovisual links in exogenous covert spatial orienting.* Perception & Psychophysics, vol. 59, no. 1, pages 1–22, 1997. (Cité en page 48.)

[Spence 2011] Charles Spence. *Crossmodal correspondences : A tutorial review.* Attention, Perception, & Psychophysics, vol. 73, no. 4, pages 971–995, 2011. (Cité en page 137.)

[Staufenbiel *et al.* 2011] Sabine M Staufenbiel, Rob H. J. van der Lubbe et Durk Talsma. *Spatially uninformative sounds increase sensitivity for visual motion change.* Experimental Brain Research, vol. 213, pages 457–464, 2011. (Cité en page 29.)

[Stein & Meredith 1993] BE Stein et MA Meredith. The Merging of the Senses. Cambridge,MA, USA, mit press édition, 1993. (Cité en pages 25, 27 et 57.)

[Stevens *et al.* 1937] Stanley Smith Stevens, John Volkman et Edwin B Newman. *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences.* Journal of the Acoustical Society of America, vol. 8, no. 3, pages 185–190, 1937. (Cité en page 123.)

[Sumby & Pollack 1954] W.H. Sumby et I. Pollack. *Visual contribution to speech intelligibility in noise*. The Journal of the Acoustical Society of America, vol. 26, pages 212–215, 1954. (Cité en page 92.)

[Talsma *et al.* 2010] Durk Talsma, Daniel Senkowski, Salvador Soto-Faraco et Marty G. Woldorff. *The multifaceted interplay between attention and multisensory integration*. Trends in Cognitive Sciences, vol. 14, no. 9, pages 400–410, 2010. (Cité en page 95.)

[Tatler *et al.* 2005] Benjamin W. Tatler, Roland J. Baddeley et Iain D. Gilchrist. *Visual correlates of fixation selection : effects of scale and time*. Vision Research, vol. 45, pages 643–659, 2005. (Cité en pages 11, 40, 47, 73 et 87.)

[Tatler *et al.* 2006] Benjamin W. Tatler, Roland J. Baddeley et Benjamin T Vincent. *The long and the short of it : Spatial statistics at fixation vary with saccade amplitude and task*. Vision Research, vol. 46, pages 1857–1862, 2006. (Cité en page 108.)

[Tatler *et al.* 2011] Benjamin W. Tatler, Mary M Hayhoe, Michael F Land et Dana H Ballard. *Eye guidance in natural vision : Reinterpreting salience*. Journal of Vision, vol. 11, no. 5, pages 1–23, 2011. (Cité en pages 18, 108 et 113.)

[Tatler 2007] Benjamin W. Tatler. *The central fixation bias in scene viewing : Selecting an optimal viewing position independently of motor biases and image feature distributions*. Journal of Vision, vol. 7, no. 14, pages 1–17, 2007. (Cité en pages 13, 47, 85 et 86.)

[Teager 1980] H. M. Teager. *Some observations on oral air flow during phonation*. IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 28, no. 5, pages 599–601, 1980. (Cité en page 50.)

[Thorpe *et al.* 1996] Simon Thorpe, Denis Fize, Catherine Marlot*et al.* *Speed of processing in the human visual system*. Nature, vol. 381, pages 520–522, 1996. (Cité en page 83.)

[Thurlow & Jack 1973] Willard R Thurlow et Charles E Jack. *Certain Determinants of the "Ventriloquism Effect"*. Perceptual and Motor Skills, vol. 36, pages 1171–1184, 1973. (Cité en page 94.)

[Tibshirnani 1996] Robert Tibshirnani. *Regression Shrinkage and Selection via the Lasso*. Journal of the Royal Statistical Society. Series B (Methodological), vol. 58, no. 1, pages 267–288, 1996. (Cité en page 76.)

[Tiippana *et al.* 2004] K. Tiippana, T S. Andersen et M. Sams. *Visual attention modulates audiovisual speech perception*. European Journal of Cognitive Psychology, vol. 16, pages 457–472, 2004. (Cité en page 94.)

[Tikhonov 1943] Andrey Nikolayevich Tikhonov. *On the stability of inverse problems*. Comptes Rendus (Doklady) de l'Academie des Sciences de l'URSS, vol. 39, no. 5, pages 195–198, 1943. (Cité en page 76.)

[Todd 1912] John Welhoff Todd. *Reaction to multiple stimuli*. Archives of Psychology, vol. 25, 1912. (Cité en page 48.)

[Torralba *et al.* 2006] Antonio Torralba, Aude Oliva, Monica S. Castelhano et John M. Henderson. *Contextual guidance of eye movements and attention in real-world scenes : The role of global features in object search.* Psychological Review, vol. 113, pages 766–786, 2006. (Cité en pages 18 et 59.)

[Tosi *et al.* 1997] Virgilio Tosi, Luciano Mecacci et Elio Pasquali. *Scanning eye movements made when viewing film : preliminary observations.* International Journal of Neuroscience, vol. 92, no. 1-2, pages 47–52, 1997. (Cité en pages 12 et 13.)

[Treisman & Gelade 1980] Anne M. Treisman et Garry Gelade. *A feature-integration theory of attention.* Cognitive psychology, vol. 12, pages 97–136, 1980. (Cité en pages 14 et 90.)

[Tseng *et al.* 2009] Po-He Tseng, Ran Carmi, Ian G M Cameron, Douglas P. Munoz et Laurent Itti. *Quantifying center bias of observers in free viewing of dynamic natural scenes.* Journal of Vision, vol. 9, no. 7, pages 1–16, 2009. (Cité en pages 13, 47 et 84.)

[Tsotsos *et al.* 2008] John K. Tsotsos, Antonio J Rodríguez-Sánchez, Albert L Rothenstein et Eugene Simine. *The different stages of visual recognition need different attentional binding strategies.* Brain Research, vol. 1225, no. C, pages 119–132, 2008. (Cité en page 18.)

[Tsuchida & Cottrell 2012] Tomoki Tsuchida et Garrison W. Cottrell. *Auditory Saliency Using Natural Statistics.* In COGSCI 2012, pages 1048–1053, Sapporo, Japan, 2012. (Cité en pages 21 et 23.)

[Unema *et al.* 2005] Pieter J. A. Unema, Pannasch Sebastian, Markus Joos et Boris M. Velichkovsky. *Time course of information processing during scene perception : The relationship between saccade amplitude and fixation duration.* Visual Cognition, vol. 12, no. 3, pages 1–22, 2005. (Cité en pages 12, 84 et 86.)

[Vajaria *et al.* 2008] Himanshu Vajaria, Sudeep Sarkar et Rangachar Kasturi. *Exploring Co-occurence between Speech and Body Movement for Audio-guided Video Localization.* IEEE Transactions on Circuits and Systems for Video Technology, vol. 18, no. 11, pages 1608–1617, 2008. (Cité en page 123.)

[Väljamäe & Soto-Faraco 2008] Aleksander Väljamäe et Salvador Soto-Faraco. *Filling-in visual motion with sounds.* Acta Psychologica, vol. 129, no. 2, pages 249–254, 2008. (Cité en page 29.)

[Van der Burg *et al.* 2008] Erik Van der Burg, Christian N L Olivers, Adelbert W Bronkhorst et Jan Theeuwes. *Pip and pop : Nonspatial auditory signals improve spatial visual search.* Journal of Experimental Psychology : Human Perception and Performance, vol. 34, no. 5, pages 1053–1065, 2008. (Cité en pages 28, 48 et 95.)

[VanRullen & Thorpe 2001] R. VanRullen et S. J. Thorpe. *Is it a bird ? Is it a plane ? Ultra-rapid visual categorisation of natural and artifactual objects.* Perception, vol. 30, pages 655–668, 2001. (Cité en page 91.)

[Vatakis & Spence 2006] Argiro Vatakis et Charles Spence. *Audiovisual synchrony perception for music, speech, and object actions.* Brain Research, vol. 1111, pages 134–142, 2006. (Cité en pages 57 et 62.)

[Vatikiotis-Bateson *et al.* 1998] Eric Vatikiotis-Bateson, Inge-Marie Eigsti, Sumio Yano et Kevin G Munhall. *Eye movement of perceivers during audiovisual-speech perception.* Perception & Psychophysics, vol. 60, no. 6, pages 926–940, 1998. (Cité en pages 97 et 109.)

[Velichkovsky *et al.* 2005] Boris M. Velichkovsky, Markus Joos, Jens R. Helmert et Sebastian Pannasch. *Two Visual Systems and their Eye Movements : Evidence from Static and Dynamic Scene Perception.* In Cognitive Science Society (XXVIIth), pages 74–86, Stresa, Italy, 2005. (Cité en pages 12 et 86.)

[Vilaró *et al.* 2012] Anna Vilaró, Andrew T Duchowski, Pilar Orero, Tom Grindinger, Stephen Tetreault et Elena di Giovanni. *How sound is the Pear Tree Story ? Testing the effect of varying audio stimuli on visual attention distribution.* Perspectives : Studies in Translatology, vol. 20, no. 1, pages 55–65, 2012. (Cité en page 62.)

[Vincent *et al.* 2009] Benjamin T Vincent, Roland J. Baddeley, Alessia Correani, Tom Troscianko et Ute Leonards. *Do we look at lights ? Using mixture modelling to distinguish between low- and high-level factors in natural image viewing.* Visual Cognition, vol. 17, no. 6-7, pages 856–879, 2009. (Cité en page 75.)

[Viola & Jones 2004] Paul Viola et Michael J Jones. *Robust real-time face detection.* International journal of computer vision, vol. 57, no. 2, pages 137–154, 2004. (Cité en page 16.)

[Võ *et al.* 2012] Melissa L. H. Võ, Tim J. Smith, Parag K. Mital et Henderson John M. *Do the eyes really have it ? Dynamic allocation of attention when viewing moving faces.* Journal of Vision, vol. 12, no. 13, pages 1–14, 2012. (Cité en pages 92, 98, 108 et 109.)

[Vroomen & de Gelder 2000] Jean Vroomen et Beatrice de Gelder. *Sound enhances visual perception : Cross-modal effects of auditory organization on vision.* Journal of experimental psychology. Human perception and performance, vol. 26, no. 5, pages 1583–1590, 2000. (Cité en page 28.)

[Vroomen & Stekelenburg 2011] Jean Vroomen et Jeroen J Stekelenburg. *Perception of intersensory synchrony in audiovisual speech : Not that special.* Cognition, vol. 118, no. 1, pages 75–83, 2011. (Cité en page 62.)

[Wagner *et al.* 2006] Philipp Wagner, Klaus Bartl, Wolfgang Günthner, Erich Schneider, Thomas Brandt et Heinz Ulbrich. *A pivotable head mounted camera system that is aligned by three-dimensional eye movements.* In Proceedings of the 2006 symposium on Eye tracking research & applications, ETRA '06, pages 117–124. ACM Press, 2006. (Cité en page 11.)

[Walker & Scott 1981] JT Walker et KJ Scott. *Auditory–visual conflicts in the perceived duration of lights, tones, and gaps.* Journal of Experimental Psycho-

logy : Human Perception and Performance, vol. 7, pages 1327–1339, 1981. (Cité en page 29.)

[Walker *et al.* 2006] Robin Walker, Eugene McSorley et Patrick Haggard. *The control of saccade trajectories : Direction of curvature depends on prior knowledge of target location and saccade latency.* Perception & Psychophysics, vol. 68, no. 1, pages 129–138, 2006. (Cité en page 47.)

[Wandell *et al.* 2007] BA Wandell, SO Dumoulin et AA Brewer. *Visual field maps in human cortex.* Neuron, vol. 56, no. 2, pages 366–383, 2007. (Cité en page 4.)

[Wang *et al.* 2010] Junle Wang, Damon M Chandler et Patrick Le Callet. *Quantifying the Relationship between Visual Salience and Visual Importance.* In Spie Human and Electronic Imaging (HVEI) XV, San Jose, USA, 2010. (Cité en page 87.)

[Wang *et al.* 2012] Helena X. Wang, Jeremy Freeman, Elisha P. Merriam, Uri Hasson et David J. Heeger. *Temporal eye movement strategies during naturalistic viewing.* Journal of Vision, vol. 12, no. 1, pages 1–27, 2012. (Cité en pages 10, 46 et 48.)

[Watanabe & Shimojo 1998] Katsumi Watanabe et Shinsuke Shimojo. *Attentional modulation in perception of visual motion events.* Perception, vol. 27, no. 9, pages 1041–1054, 1998. (Cité en page 29.)

[Welch & Warren 1980] R. B. Welch et D. H. Warren. *Immediate perceptual response to intersensory discrepancy.* Psychological Bulletin, vol. 88, pages 638–667, 1980. (Cité en page 93.)

[Welch *et al.* 1986] RB Welch, LD DuttonHurt et DH Warren. *Contributions of audition and vision to temporal rate perception.* Perception & Psychophysics, vol. 39, pages 294–300, 1986. (Cité en page 29.)

[Wolfe 1994] Jeremy M Wolfe. *Guided Search 2.0 - A revised model of visual search.* Psychonomic Bulletin & Review, vol. 1, no. 2, pages 202–238, 1994. (Cité en page 90.)

[Wright *et al.* 2010] John Wright, Yi Ma, Julien Mairal, Guillermo Sapiro, Thomas S Huang et Shuicheng Yan. *Sparse representation for computer vision and pattern recognition.* Proceedings of the IEEE, vol. 98, no. 6, pages 1031–1044, 2010. (Cité en page 76.)

[Wu *et al.* 2010] Chia-Chien Wu, Oh-Sang Kwon et Eileen Kowler. *Fitts's Law and speed/accuracy trade-offs during sequences of saccades : Implications for strategies of saccadic planning.* Vision Research, vol. 50, no. 21, pages 2142–2157, 2010. (Cité en page 47.)

[Xu *et al.* 2014] Juan Xu, Ming Jiang, Shuo Wang, Mohan S Kankanhalli et Qi Zhao. *Predicting human gaze beyond pixels.* Journal of Vision, vol. 14, no. 1, pages 1–20, 2014. (Cité en page 18.)

[Yang *et al.* 2002] Qing Yang, Maria Pia Bucci et Zoï Kapoula. *The Latency of Saccades, Vergence, and Combined Eye Movements in Children and in Adults.* Investigative Ophthalmology & Visual Science, vol. 43, no. 9, pages 2939–2949, 2002. (Cité en page 47.)

[Yarbus 1967] Alfred. L. Yarbus. Eye Movements and Vision. Plenum, New York, USA, 1967. (Cité en pages 7, 64, 85 et 90.)

[Yi & Xu 2008] Nengjun Yi et Shizhong Xu. *Bayesian LASSO for quantitative trait loci mapping.* Genetics, vol. 179, no. 2, pages 1045–1055, 2008. (Cité en page 76.)

[Young 1994] S.J. Young. *The HTK Hidden Markov Model Toolkit : Design and Philosophy.* Entropic Cambridge Research Laboratory, Ltd, 1994. (Cité en page 123.)

[Zaraki *et al.* 2014] Abolfazl Zaraki, Daniele Mazzei, Manuel Giuliani et Danilo De Rossi. *Designing and Evaluating a Social Gaze-Control System for a Humanoid Robot.* IEEE Transactions on Human-Machine Systems, vol. 44, no. 2, pages 157–168, 2014. (Cité en pages 21 et 33.)

[Zhang *et al.* 2008] Lingyun Zhang, Matthew H Tong, Tim K Marks, Honghao Shan et Garrison W. Cottrell. *SUN : A Bayesian framework for saliency using natural statistics.* Journal of Vision, vol. 8, no. 7, pages 1–20, 2008. (Cité en page 59.)

[Zhao & Koch 2012] Qi Zhao et Christof Koch. *Learning visual saliency by combining feature maps in a nonlinear manner using AdaBoost.* Journal of Vision, vol. 12, no. 6, pages 1–15, 2012. (Cité en page 16.)

[Zlatintsi *et al.* 2012] Athanasia Zlatintsi, Petros Maragos, Alexandros Potamianos et Georgios Evangelopoulos. *A Saliency-Based Approach to Audio Event Detection and Summarization.* In European Signal Processing Conference (EUSIPCO 2012), pages 1294–1298, Bucharest, Romania, 2012. (Cité en pages 21 et 32.)

Annexe A

Stimuli de l'experience 1

Vidéos	Durée (s)	Contenu Visuel	Contenu Sonore
1	21.8	objets divers	bruits mécaniques
2	57.2	bourdons butinant	bourdonnement
3	26.8	phoques sur banquise	bruits aquatiques
4	57.2	phoques et crabes	mer et cris de phoques
5	31.7	habitants des fonds marins	bruits de sable
6	41.7	tempête	vent et vagues
7	9.6	avions de chasse	bruit des avions
8	19.4	canards	musique entraînante
9	65.2	troupeau de cerfs	musique d'ambiance
10	21	banc de dauphins	musique rapide
11	58.5	tortues et oiseaux	musique dramatique
12	32.4	intérieur de voiture	voix off
13	29.7	chutes d'eau	voix off
14	21.2	poissons	voix off
15	24.8	poissons	voix off
16	34.6	terre vue du ciel	voix off
17	28.4	paysage	calme
18	18.4	frondaison	gazouillis
19	14.2	hautes herbes	cigales
20	23.4	frondaison	gazouillis
21	19.4	frondaison	vent dans les feuilles
22	26.4	sous bois	gazouillis
23	28.7	montagne	flûte andine
24	20.5	rizière	percussion métal

suite page suivante

suite de la page précédente

25	8.1	désert	musique entraînante
26	25.1	iceberg	violons
27	15.9	forêt survolée	chants choraux
28	15	forêt + montagne	musique calme
29	22.8	glacier	voix off
30	27.9	lacs	voix off + musique
31	31.8	terre vue de l'espace	voix off
32	27.4	bocage + brouillard	voix off
33	33.4	alpage	voix off radio
34	24.7	femme assise + passagers	bruits du metro
35	33.4	homme se rapprochant	bruit de verre
36	32.9	homme dans train	bruit du train
37	7.7	3 adolescents	mobylette
38	17.6	2 personnes attablées	vaisselle
39	20.9	2 personnes dans la rue	bruits de pas
40	30.5	indiens dans le Gange	musique d'ambiance
41	35.2	indiens à vélo	musique
42	19.4	2 personnes assises	flamenco
43	24.9	homme dansant	musique entraînante
44	15.8	femme + homme	musique calme
45	22.9	homme titubant	musique de cirque
46	22.3	2 actrices théâtre	dialogue
47	53.2	homme et femme assis	chanson
48	19.7	2 enfants	dialogue
49	23.6	4 hommes	dialogue
50	29.5	2 hommes	dialogue

Table A.1 – Description de la base de vidéos utilisée lors de l'expérience 1.

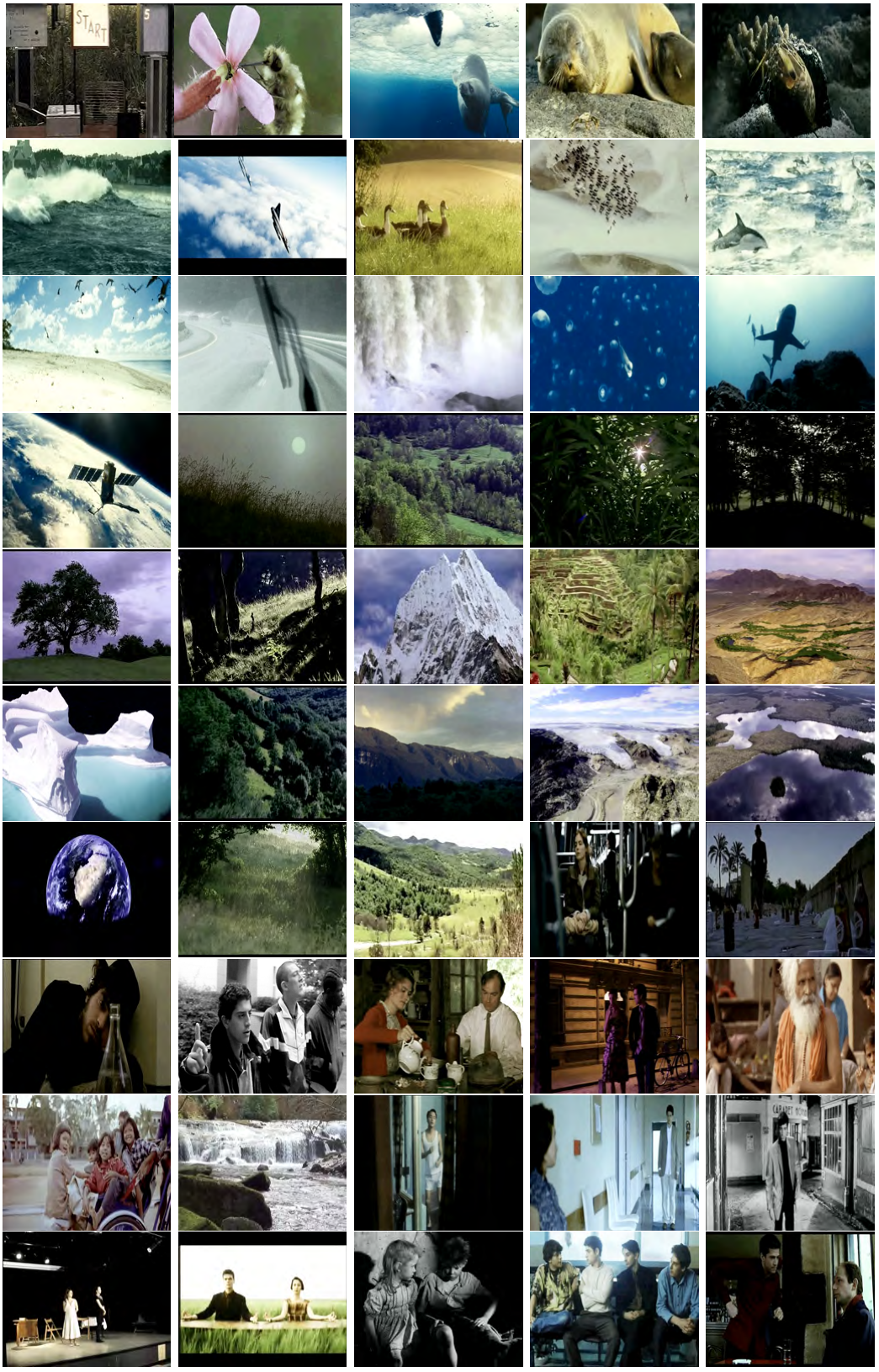


Figure A.1 – Frames extraites des vidéos 1 à 50 (de haut en bas et de gauche à droite) utilisées lors de l'expérience 1.

Annexe B

Stimuli de l'experience 2

Vidéos	Durée (s)	Contenu Visuel	Contenu Sonore
1	24.3	voiture et pelleteuse	bruits métalliques
2	14.9	voiture et pelleteuse	bruits métalliques
3	14.7	château d'eau et machine	bruits du marteau
4	15.5	cloche	cloche
5	16.1	mur et pelleteuse	bruits de démolition
6	15.4	hélicoptère téléguidé	bruit des pales
7	10.6	voiture dans décharge	bruit moteur
8	8.7	robot sauteur	impact du robot
9	12.1	marteau sur pierre	marteau + pierre cassée
10	24	avion téléguidé	bruit de l'avion
11	9.3	voiture de police	sirène
12	11.9	voiture de rallye	moteur
13	22.7	camion pompier	sirène
14	14.9	neige tombe du toit	chute de neige
15	15.5	robot soudeur	soudure
16	23.8	horloge mécanique	carillon et mécanismes
17	15.3	flipper en lego	bille, roues, leviers
18	11.9	fontaine remplie bambou	eau et bambou
19	10.9	circuit lego	bille, sirène
20	22.7	machine à café	bruits de la machine, jet de billes
21	19.1	réaction en chaine	bruits des constituants
22	12.9	réaction en chaine	bruits des constituants
23	21.5	réaction en chaine	bruits des constituants
24	14.9	lavabo douche	eau et objets tombants

suite page suivante

suite de la page précédente

25	14.2	couverts, balles	objets tombants
26	10.4	fenêtre, rue	store
27	16.4	cuisine	machine à laver, objets divers
28	21.9	mécanisme automatique	billes
29	9.5	oies + bateau	piaillement + sirène
30	23.8	table	objets tombants, téléphone
31	15.5	ruisseau	eau
32	14.4	clôture	vent
33	13.9	lisière de forêt	vent
34	21.1	nuages	vent
35	13.9	ruisseau	eau
36	19.1	ruisseau	eau
37	9.9	prairie et brume	vent
38	9.9	marécage et brume	vent
39	19.9	bord de mer	vagues
40	14.9	lac	clapotis
41	14.2	pont enneigé	neige fondue
42	19.4	chute d'eau	eau
43	15.5	prairie	vent
44	18.9	bord de mer	vagues
45	14.9	champ et nuages	vent
46	21.5	2 personnes	dialogue
47	18.9	3 personnes	dialogue
48	13.7	2 personnes	dialogue
49	17.5	2 personnes	dialogue
50	26.6	2 personnes	dialogue
51	21.1	2 personnes	dialogue
52	11.9	2 personnes	dialogue
53	19.7	2 personnes	dialogue
54	23.9	2 personnes	dialogue
55	21.4	2 personnes	dialogue

suite page suivante

suite de la page précédente

56	23.6	4 personnes	dialogue
57	29.5	2 personnes	dialogue
58	16.8	2 personnes	dialogue
59	15.2	2 personnes	dialogue
60	12.3	2 personnes	dialogue

Table B.1 – Description de la base de vidéos utilisée lors de l'expérience 2. Les vidéos de 1 à 15 correspondent à la catégorie Un Objet en Mouvement (UOM), de 16 à 30 à la catégorie Plusieurs Objets en Mouvement (POM), de 31 à 45 à la catégorie Paysages et de 45 à 60 à la catégorie Visages. Les durées indiquées correspondent à celle des vidéos dans la condition sonore Originale. Dans les autres conditions, ces durées peuvent différer de quelques secondes afin d'être de la même longueur que la bande-son associée.

Figure B.1 – Frames extraites des vidéos de la catégorie visuelle Un Objet en Mouvement (UOM) de l'expérience 2. Elles correspondent aux vidéos 1 à 15.

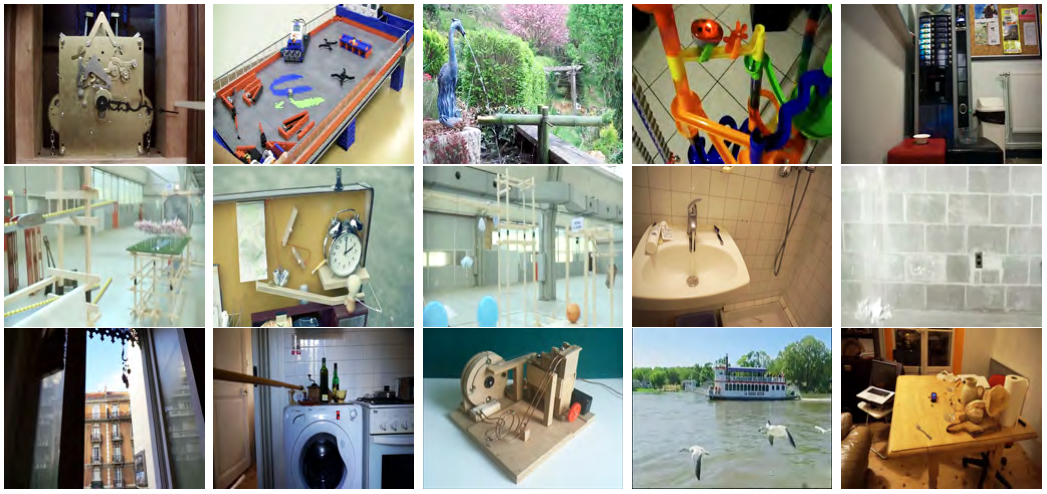


Figure B.2 – Frames extraites des vidéos de la catégorie visuelle Plusieurs Objets en Mouvement (POM) de l'expérience 2. Elles correspondent aux vidéos 16 à 30.

Figure B.3 – Frames extraites des vidéos de la catégorie visuelle Paysages de l'expérience 2. Elles correspondent aux vidéos 31 à 45.

Figure B.4 – Frames extraites des vidéos de la catégorie visuelle Visages de l'expérience 2. Elles correspondent aux vidéos 46 à 60.

Annexe C

Détail des ANOVA du chapitre 3

Cette annexe reprend certaines des ANOVA effectuées sur les résultats de l'expérience 2. Pour chaque catégorie visuelle (Visages, Paysages, Un Objet en Mouvement (UOM), Plusieurs Objets en Mouvement (POM)), nous détaillons l'effet des conditions expérimentales (Originale, Son Visages, Son Paysages, Mix Intra, Son POM, Son UOM) sur **(1)** les différentes métriques utilisées dans ce chapitre, et **(2)** les poids estimés par modélisation statistique.

C.1 Métriques

Cette section reprend et détaille les résultats des ANOVA présentées à la section 3.2.2.2. Pour chaque catégorie visuelle et chaque métrique, nous avons mené une ANOVA à un facteur intra (les conditions expérimentales). Pour la dispersion et la distance au centre, chaque niveau a 15 items (les stimuli). Pour les amplitudes de saccade et les durées de fixation, chaque niveau a 72 items (les participants).

Visages

Dispersion Il existe un effet principal de la condition expérimentale ($F(3,42) = 17.97$, $p < .001$). Des comparaisons a posteriori de Bonferroni montrent que la dispersion dans la condition Originale est inférieure à toutes les autres (tous les $p < .001$). Il n'existe pas de différence significative entre les autres conditions expérimentales (Son POM vs. Mix Intra, $p = .07$; Son Paysages vs. Son POM, $p = 1$; Son Paysages vs. Mix Intra, $p = .79$).
Distance au centre Il n'y a pas d'effet de la condition expérimentale ($F(3,42) = .38$, $p = .77$).
Amplitude de saccade Il existe un effet principal de la condition expérimentale ($F(3,213) = 6.2$, $p < .001$). Des comparaisons a posteriori de Bonferroni montrent que les moyennes des amplitudes de saccade médianes dans la condition Originale sont inférieures à toutes les autres (tous les $p < .01$). Il n'existe pas de différence significative entre les autres conditions expérimentales (tous les $p = 1$).
Durée de fixation Il n'y a pas d'effet de la condition expérimentale ($F(3,213) =$

.44, $p = .73$).

Paysages

Dispersion Il n'y a pas d'effet de la condition expérimentale ($F(3,42) = .83$, $p = .49$).
Distance au centre Il n'y a pas d'effet de la condition expérimentale ($F(3,42) = 1.2$, $p = .32$).
Amplitude de saccade Il n'y a pas d'effet de la condition expérimentale ($F(3,213) = .22$, $p = .88$).
Durée de fixation Il n'y a pas d'effet de la condition expérimentale ($F(3,213) = 1.57$, $p = .2$).

Un Objet en Mouvement

Dispersion Il n'y a pas d'effet de la condition expérimentale ($F(3,42) = .20$, $p = .9$).
Distance au centre Il n'y a pas d'effet de la condition expérimentale ($F(3,42) = 1.2$, $p = .34$).
Amplitude de saccade Il n'y a pas d'effet de la condition expérimentale ($F(3,213) = 1.6$, $p = .19$).
Durée de fixation Il n'y a pas d'effet de la condition expérimentale ($F(3,213) = .40$, $p = .75$).

Plusieurs Objets en Mouvement

Dispersion Il n'y a pas d'effet de la condition expérimentale ($F(3,42) = .16$, $p = .93$).
Distance au centre Il n'y a pas d'effet de la condition expérimentale ($F(3,42) = .76$, $p = .53$).
Amplitude de saccade Il n'y a pas d'effet de la condition expérimentale ($F(3,213) = .32$, $p = .81$).
Durée de fixation Il n'y a pas d'effet de la condition expérimentale ($F(3,213) = .20$, $p = .89$).

C.2 Modélisation statistique

Cette section reprend les résultats des ANOVA sur les poids des attributs visuels (saillance statique, saillance dynamique, biais de centralité et carte uniforme) estimés par Espérance - Maximisation (EM) et Least Absolute Shrinkage and Selection

Operator (Lasso). Ces résultats ont été présentées à la section 3.3.2. Pour chaque catégorie visuelle, nous avons mené une ANOVA à deux facteurs intra (les attributs visuels et les conditions expérimentales). Chaque niveau a 15 items (les stimuli). Comme les poids des attributs visuels ont déjà été comparés section 3.3.2, nous nous bornons ici à décrire les effets des conditions expérimentales ainsi que leurs interactions.

C.2.1 Estimation Espérance - Maximisation

Visages

Il n'y a pas d'effet des conditions expérimentales ($F(3,42) = .90$, $p = .45$) ni de leur interaction avec les attributs visuels ($F(9,126) = .08$, $p = .99$).

Paysages

Il n'y a pas d'effet des conditions expérimentales ($F(3,42) = 1.26$, $p = .30$) ni de leur interaction avec les attributs visuels ($F(9,126) = .28$, $p = .98$).

Un Objet en Mouvement

Il n'y a pas d'effet des conditions expérimentales ($F(3,42) = .83$, $p = .48$) ni de leur interaction avec les attributs visuels ($F(9,126) = .06$, $p = .99$).

Plusieurs Objets en Mouvement

Il n'y a pas d'effet des conditions expérimentales ($F(3,42) = .89$, $p = .45$) ni de leur interaction avec les attributs visuels ($F(9,126) = .03$, $p = 1$).

C.2.2 Estimation Lasso

Visages

Il n'y a pas d'effet des conditions expérimentales ($F(3,42) = 1.14$, $p = .34$) ni de leur interaction avec les attributs visuels ($F(9,126) = .02$, $p = 1$).

Paysages

Il n'y a pas d'effet des conditions expérimentales ($F(3,42) = 1.28$, $p = .29$) ni de leur interaction avec les attributs visuels ($F(9,126) = .33$, $p = .96$).

Un Objet en Mouvement

Il n'y a pas d'effet des conditions expérimentales ($F(3,42) = .41$, $p = .74$) ni de leur interaction avec les attributs visuels ($F(9,126) = .03$, $p = 1$).

Plusieurs Objets en Mouvement

Il n'y a pas d'effet des conditions expérimentales ($F(3,42) = .54$, $p = .66$) ni de leur interaction avec les attributs visuels ($F(9,126) = .07$, $p = .99$).

Annexe D

Algorithme d'Espérance-Maximisation

Data: $X(x,t)_{k\in[1..p]}$ les densités de probabilité de chaque variable du modèle, et les positions oculaires de chaque participant.
Result: $\beta_k(t)$ les poids de chaque variable ($\sum_{k=1}^{p}\beta_k(t)=1$)

for *chaque frame* $f = 1..nframe$ **do**
  Initialization : $\forall k \in [1..p], \beta_k^{(0)}(f) = \frac{1}{p}$
    iteration m = 0
    DiffLog=1 ;
  while *DiffLog* $\geq$ *threshold* **do**
    m = m + 1;
    // Espérance - soit l'estimation courante des $\beta_k^{(m-1)}$
    for *chaque participant* $s = 1..nsub$ **do**
      x_s^f = position oculaire du participant s sur la frame f ;
      for *chaque variable* $k = 1..p$ **do**
        Estim(s,k) = $\beta_k^{(m-1)}(f).X_k(x_s^f, f)$;
      end
    end
    // Maximisation - mise à jour des poids $\beta_k^{(m-1)}$ à $\beta_k^{(m)}$
    for *chaque variable* $k = 1..p$ **do**
      for *chaque participant* $s = 1..nsub$ **do**
        Maxim(s,k) = $\frac{\text{Estim(s,k)}}{\sum_{k=1}^{p}\text{Estim(s,k)}}$;
      end
      $\beta_k^{(m)}(f) = \frac{\sum_{s=1}^{nsub}\text{Maxim(s,k)}}{nsub}$;
    end
    LogLikelihood = $\frac{\sum_{s=1}^{nsub}\log\sum_{k=1}^{p}\text{Estim(s,k)}}{nsub}$;
    DiffLog = abs(LogLikelihood$^{(m)}$ - LogLikelihood$^{(m-1)}$);
  end
end

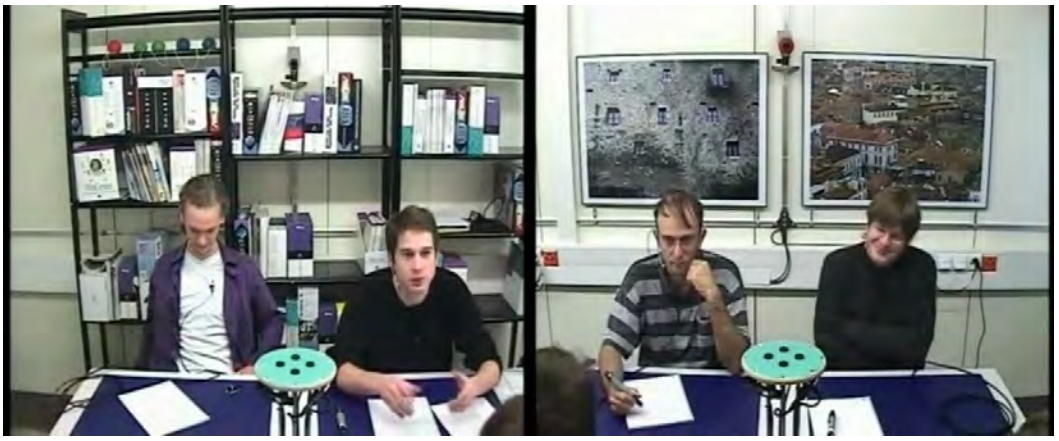
Algorithm 1: Algorithme d'Espérance-Maximisation (EM) utilisé section 3.3.1.1.

Annexe E

Stimuli de l'experience 3

	durée (s)	Changements locuteur	Temps de parole (%)			
			vis. 1	vis. 2	vis. 3	vis. 4
IN1008_1	35	3	42	41	17	0
IN1008_2	45	3	0	60	5	35
IN1008_3	80	4	0	32	6	60
IN1008_4	22	1	0	67	0	33
IN1008_5	35	1	0	53	0	44
IN1012_1	20	2	0	0	85	14
IN1012_2	60	3	72	24	0	4
IN1012_3	55	1	39	60	0	0
IN1012_4	60	5	45	52	0	3
IN1012_5	25	4	91	28	0	0
IN1014_1	50	1	0	55	45	0
IN1014_2	59	9	0	51	0	48
IN1014_3	40	4	0	13	31	56
IN1014_4	60	3	22	1	76	1
IN1014_5	40	2	0	3	67	33

Table E.1 – Durée, nombre de tours de parole et temps de parole de chacun des locuteurs présents dans les stimuli utilisés lors de l'expérience 3. Les visages sont numérotés de la gauche vers la droite. Si la somme des temps de parole est inférieure à 100%, c'est qu'il y a des moments de silence.



(a) réunion IN1008

(b) réunion IN1012

(c) réunion IN1014

Figure E.1 – Frames extraites des trois réunions de travail utilisées pour créer les stimuli de l'expérience 3.

Annexe F

Curriculum Vitæ

Antoine Coutrot

PhD Student

3 rue Marceau
38000 Grenoble, France
+33 (0)6 87 64 50 98
acoutrot@gmail.com
27 years old, French

Research Interests

The key to robust perception is the efficient combination and integration of multiple sources of sensory information. My research focuses on multimodal perception of naturalistic scenes. I follow a dual approach. First, I collect behavioral data through eye-tracking experiments during which participants watch videos in different audio-visual conditions. Then I interpret the results using statistical modeling. I aim at (1) understanding the mechanisms of cross-modal integration and (2) modeling them to improve attention models.

Education

2011–2014 **PhD Student**, *Gipsa-lab, Vision and Brain Signal Processing Team*, Grenoble, France.

2010–2011 **M. Sc**, *Cognitive Science*, Grenoble University, France.

2007–2011 **Engineer school**, *Grenoble Institute of Technology (Phelma)*, Grenoble-INP, France.

2005–2007 **Preparatory Classes, Mathematics and Physics**, *MPSI–MP**, Paris, France.

Professional activities

PhD

PhD subject *How does sound impact on visual exploration of videos ? Behavioural and computational approaches - Integration of auditory information in visual saliency models.*

laboratory Gipsa-lab - Joint Research Unit CNRS and Grenoble University

supervisors Prof. Alice Caplier & Dr. Nathalie Guyader

funding doctoral contract funded by the Ministry for Higher Education and Research

Teaching

since 2012 **Lectures and tutorials**, *Grenoble Institute of Technology*, Grenoble, France, Level: Engineer School - from BSc (3rd year) to MSc.

- Mathematics: Lectures (25h) and Tutorials (48h)
  Complex Analysis, Partial differential equations, Fourier and Laplace transforms
- Signal Processing: Tutorials (55h)
  Digital Signal Processing, Spectral Analysis, Speech Processing, Image Processing
- Master thesis advised
  Philippe Benjamin - Implementation of a Graphical User Interface for Eye-Tracking Data, 4 months (2013)
- Master thesis Jury committees (9h)

Internships

2011 (5 months) **M.Sc**, *Gipsa-lab*, Vision and Brain Signal Processing Team, Grenoble University, France.
subject Impact of sound on eye movements during free exploration of dynamic natural scenes.
supervisors Nathalie Guyader, Bertrand Rivet & Gelu Ionescu.
2009 (4 months) **Junior Engineer**, *Petzl (technical equipment for mountain sports)*, Crolles, France.
subject Aging simulation for textile products, material strength (supervisor: Carole Dubois).

Skills

Multimodal Perception

- Eye-tracking experiments with static and dynamic stimuli (Eyelink 1000)
- Eye-movements analysis: scanpath, clustering, divergence...

Audiovisual Signal Processing

- Visual and auditory saliency models (design and validation)
- Cross-modal analysis, Mutual Information

Statistical analysis and modeling

- Linear Methods for Regression and Classification: subset selection and shrinkage (Least Squares, Lasso)
- Model Inference and Averaging (Expectation-Maximization)
- Model Assessment and Selection: Bayesian Approach (BIC)
- Unsupervised Learning: Cluster Analysis (K-means, Mean Shift, Gaussian Mixtures)

Computing

Operating Systems	GNU/Linux, Mac OS, Microsoft Windows	Langages	**Matlab** (extensive experience), R, HTML
Scientific Tools	Statistica, Maple, Mathematica	Desktop publishing	LaTeX, Office, Adobe, QuarkXPress

Languages

French	**mother tongue**	
English	**fluent**	*TOEFL*
Spanish	**conversational**	*6 months experience in Argentina*
Russian	**basic**	*high school*

Publications

Journal Papers

- Coutrot, Antoine and Guyader, Nathalie. *How Saliency, Faces and Sound influence gaze in Dynamic Social Scenes*, Journal of Vision, Vol. 14, No. 8, pp 1-17, 2014.
- Coutrot, Antoine and Guyader, Nathalie and Ionescu, Gelu and Caplier, Alice. *Video viewing: do auditory salient events capture visual attention?*, Annals of Telecommunications, Vol. 69, No. 1 pp 89-97, 2014.
- Coutrot, Antoine and Guyader, Nathalie and Ionescu, Gelu and Caplier, Alice. *Influence of soundtrack on eye movements during video exploration*, Journal of Eye Movement Research, Vol. 5, No. 4, pp 1-10, 2012.

International Conferences

- Coutrot, Antoine and Guyader, Nathalie. *An audiovisual attention model for natural conversation scenes*, IEEE International Conference on Image Processing (ICIP 2014), October 2014, Paris, France.
- Coutrot, Antoine and Guyader, Nathalie. *Exploration of dynamic natural scenes: influence of unrelated soundtracks on eye movements*, In Proc. 17th European Conference on Eye Movements (ECEM 2013), August 2013, Lund, Sweden.
- Coutrot, Antoine and Guyader, Nathalie. *Toward the Introduction of Auditory Information in Dynamic Visual Attention Models*, In Proc. 14th International Workshop on Image and Audio Analysis for Multimedia Interactive Services (WIAMIS 2013), July 2013, Paris, France. **Best paper award.**

National Conferences

- Coutrot, Antoine and Guyader, Nathalie. *Intégration de l'information sonore et visuelle pour prédire les régions regardées dans des scènes de conversations*, In La Saillance Visuelle : de son exploitation à son évaluation, Journée transverse GdR ISIS et GdR Vision, June 2014, Paris, France.
- Coutrot, Antoine and Guyader, Nathalie. *Modèle de saillance audiovisuelle pour des scènes de conversations dynamiques*, In Workshop Eye-Tracking Regard Interactions et Suppléances (ERIS 2014), June 2014, Paris, France.
- Coutrot, Antoine and Caplier, Alice and Guyader, Nathalie. *Look at me when I'm talking! Influence of sound on visual exploration of conversation scenes*, EEATS Doctoral School PhD Students' Day, April 2014, Grenoble, France. **Best presentation award.**
- Coutrot, Antoine and Guyader, Nathalie. *How sound influences the visual exploration of dynamic scenes*, In 7th Annual Forum of GDR-Vision, October 2013, Paris, France.
- Coutrot, Antoine and Guyader, Nathalie. *Prise en compte du son dans les modèles d'attention visuelle dynamiques*, 4th Meeting of the Grenoble Cognition Pole, June 2013, Grenoble, France.
- Coutrot, Antoine and Caplier, Alice and Guyader, Nathalie. *Impact of Sound on Visual Exploration*, EEATS Doctoral School PhD Students' Day, March 2013, Grenoble, France. **Best presentation award.**
- Coutrot, Antoine and Guyader, Nathalie and Ionescu, Gelu and Caplier, Alice. *Exploration libre de vidéos : influence du son sur les mouvements oculaires consécutifs à un événement sonore saillant*, Actes des 15èmes journées COmpression et REprésentation des Signaux Audiovisuels (CORESA 2012), May 2012, Lille, France. **Best paper award.**

- Coutrot, Antoine and Caplier, Alice and Guyader, Nathalie. *Influence of soundtrack on eye movements*, EEATS Doctoral School PhD Students' Day, March 2012, Grenoble, France. **Best presentation award.**

Miscellaneous

Responsibilities

since 2013 **Elected representative of the Ph.D. students at the Graduate School concil**, *EEATS Graduate School, Grenoble University, France*.

Sabbatical year

H1 2010 **Volunteer in a day-care center**, *Assistant teacher with 1-5 years old children*, Bahia Blanca, Argentina.

H2 2009 **Volunteer in a sailing school (les Glénans)**, *Assistant monitor, ships maintenance*, Arz island, Morbihan, France.

Interests

Mountain	climbing, mountaineering, ski touring
Sailing	liveaboard
Culture	reading, cinema, drama, music, journalism

References

- **Dr. Nathalie Guyader, PhD**
  Dr. Nathalie Guyader is my PhD co-advisor.
  Gipsa-lab, CNRS & Grenoble University
  Vision and Brain Signal Processing team
  11 rue des mathématiques, 38402 Grenoble Campus, France
  nathalie.guyader@gipsa-lab.fr
  +33 (0)4 76 57 43 72

- **Prof. Alice Caplier, PhD**
  Prof. Alice Caplier is my PhD co-advisor.
  Gipsa-lab, CNRS & Grenoble University
  Architecture, Geometry, Perception, Images and Gestures team
  11 rue des mathématiques, 38402 Grenoble Campus, France
  alice.caplier@gipsa-lab.fr
  +33 (0)4 76 57 43 63

- **Prof. Jean-Luc Schwartz, PhD**
  Prof. Jean-Luc Schwartz was my teacher and is a collaborator.
  Gipsa-lab, CNRS & Grenoble University
  Speech, Multimodality and Development team
  11 rue des mathématiques, 38402 Grenoble Campus, France
  jean-luc.schwartz@gipsa-lab.fr
  +33 (0)4 76 57 47 12

Résumé — Nous étudions l'influence de différents attributs audiovisuels sur l'exploration visuelle de scènes naturelles dynamiques. Nous démontrons que si la façon dont nous explorons une scène dépend avant tout de son contenu visuel, dans certaines situations le son influence significativement les mouvements oculaires. La présence de son assure une meilleure cohérence entre les positions oculaires de différents observateurs, attirant leur attention et donc leur regard vers les mêmes régions. L'effet du son se retrouve tout particulièrement dans les scènes de conversation, où la présence du signal de parole associé augmente le nombre de fixations sur le visage des locuteurs, et donc la cohérence entre les scanpaths. Nous proposons un modèle de saillance audiovisuelle repérant automatiquement le visage des locuteurs afin d'en rehausser la saillance. Ces résultats s'appuient sur les mouvements oculaires de 148 participants enregistrés sur un total de plus de 75 400 frames (125 vidéos) dans 5 conditions expérimentales différentes.

Mots clés : mouvements oculaires, scènes naturelles dynamiques, saillance audiovisuelle, intégration, visages.

Abstract — We study the influence of different audiovisual features on the visual exploration of dynamic natural scenes. We show that, whilst the way a person explores a scene primarily relies on its visual content, sound sometimes significantly influences eye movements. Sound assures a better coherence between the eye positions of different observers, attracting their attention and thus their gaze toward the same regions. The effect of sound is particularly strong in conversation scenes, where the related speech signal boosts the number of fixations on speakers' faces, and thus increases the consistency between scanpaths. We propose an audiovisual saliency model able to automatically locate speakers' faces so as to enhance their saliency. These results are based on the eye movements of 148 participants recorded on more than 75,400 frames (125 videos) in 5 different experimental conditions.

Keywords : eye movements, natural dynamic scenes, audiovisual saliency, integration, faces.

Laboratoire Grenoble Image Parole Signal Automatique (Gipsa-lab)
Equipe Vision and Brain Signal Processing
Grenoble, France